

Nature's 2001

This publication is set to be ever more proactive in its development of content and services. But reflections are appropriate as we finally enter a new millennium, as well as some statements of commitment.

Nature's primary aim every week, since our beginnings in 1869, has been and will be to keep researchers and society at large in touch with the state of the best of scientific research and the people who do it. One important secondary aim, to keep finding ways to do that better, will also continue unabated.

In achieving those goals, *Nature's* independence of any scientific organization remains one of our most important assets. We believe it maximizes our ability to respond rapidly to new developments and opportunities, and also that it guarantees a precious freedom to publish heterodox views — our own or others' — in an unrestrained but responsible manner.

Another asset, and a corollary of our independence, is our internationalism. Although *Nature's* content strongly reflects the United States' great and increasing strength in the world of science, no one could complain at the high proportion of our authors from, and of pages devoted to, other countries. Our continuing commitment to that balancing act is reflected by our organizing of a conference last month on Central and Eastern Europe while also deciding to open, this month, an editorial office in San Francisco. In addition, we are playing a leading role in a collaborative project developing free online facilities for scientists in the developing world. And our circulation, which has grown continuously by about 20% to 64,000 over the past four years and shows no sign of reaching a plateau, reflects this internationalism and regional strength, with 50% in the United States.

Nature has recently grown significantly more proactive in its commissioning of outstanding authors. Launched in the middle of last year, the series of monthly 'Insight' collections of review articles has boosted enormously the amount of authoritative overview we deliver to readers, while the number of traditional review articles has also grown. And in a more forward-looking spirit, later this month we will publish an Insight survey of future science and technologies, complementing our supplement *Impacts of Foreseeable Science* published in December 1999.

Innovative writing

We will also build on another of last year's innovations: the launch of weekly News Features, greatly expanding our coverage of the research community as it responds to key scientific developments. This aspect of editorial proactivity is set to increase as more investment in top-quality journalism enhances our provision yet further over the next few months.

Over the past year or two, essays in *Nature* have provided a welcome complement to the traditional fare of comment, papers and news. To judge by readers' responses, the series of Millennium Essays has well fulfilled our original hopes of providing stimulating accounts of little-known gems of history. The Futures series — essays too, albeit fictional — has entertained in its own unorthodox way, right up to the final example at the end of last year. The Millennium series will end later this month, and will be immediately followed by the launch of two new series. (Essay series in *Nature* were initiated by Martin Kemp's on art, image and science, now collected and embellished in a book entitled *Visualisations: The Nature Book of Art and Science*, published by Oxford

University Press and University of California Press, Berkeley.)

Our essays have also had the merit of being widely readable — something not to be taken for granted in a scientific periodical. Indeed, readability is something *Nature* has cultivated systematically over the past few years. Although there are limits to what can be achieved in a concise space with interdisciplinary communication, we are committed to enhancing yet further the readability of our more technical sections, including our original papers. (Readers may not recognize the enhancements in readability that we inject into our papers' introductions, but believe us, things could be much, much worse.) And readability is not only about words. The comprehensibility and impact of much of our content have been greatly enhanced by imaginative design and improved illustrations.

Service cultures

Other areas of investment have been in editorial services to authors. Following the introduction of electronic submission last year, the proportion of authors using that route grew rapidly and dramatically to a considerable majority. We have extended electronic refereeing and communication to all manuscripts, whether submitted electronically or not. That, combined with the developing ability to publish selected papers promptly on the web, is speeding up our publication times significantly. Some papers of particularly great public interest have been made freely available on our website ahead of print publication. Online readers will also have noticed improvements to our website within the past month — of which there are more to come.

The quality of research that we publish continues to be paramount, and the resources devoted to the selection process continue to grow. In that regard, although impact factors are by no means everything, we do enjoy the fact that ours is the highest among the general science journals.

Competitors are always with us and help keep us on our mettle. The newest — but still unresolved — agents of potential change are those who make published papers freely available on the web. Anyone who has examined those activities will realize that at least some of those publishers have not yet achieved a sustainable financial model. For our part, *Nature* and our publishers stand by a policy of reasonable pricing for high quality content and services. We stand by the value we add all through the publishing process, and the professionalism and responsiveness that a well resourced and dedicated editorial team can bring. However, we are examining ideas of free access as much as an opportunity as they are a challenge.

Questions about access to original research are part of a wider debate about the ethics, politics and economics of public and private ownership in science in all its aspects. This contentious set of issues will be affected by decisions in society as well as the science community: decisions about protection of intellectual property and of data, but also about a public and political resolve in further increasing the public funding of research, amid burgeoning scientific opportunities and related societal controversies. One way or another, publishers are not the only participants in these debates facing challenges to their assumptions in the near future. ■





Most wanted
Computer modellers
draw up hit list for
protein structures
p4



Atomic legacy
Children of bomb
survivors to join in
genetic study
p5



Green light
UK parliament gives
go-ahead to work on
embryonic stem cells
p5



News feature:
superconductivity
Trying to explain a
remarkable effect
p8

Ruling on cleaner diesel leaves Republicans facing a dilemma

Tony Reichhardt, Washington

The US Environmental Protection Agency (EPA) has issued strict new regulations governing diesel exhaust from vehicles. The move has been hailed as the most significant single action to improve US air quality since lead in gasoline was banned in 1986.

The regulations — issued late last month by the outgoing administration of President Bill Clinton — follow years of feuding between academic and industrial scientists over whether small particles in diesel exhaust are hazardous to human health.

And, observers say, they throw down a gauntlet for Christine Todd Whitman, the governor of New Jersey, who will head the EPA under George W. Bush's Republican administration. Whitman will have to decide whether to keep the rules, and anger some of Bush's industrial supporters, or to resume the arduous scientific advisory process needed to draw up an alternative plan.

The new regulation requires diesel trucks and buses to be fitted with pollution control devices similar to catalytic converters. Because high-sulphur fuels choke the devices, the EPA also ordered a 97% reduction in the sulphur content of diesel fuel, to 15 parts per million. Vehicle manufacturers must meet the limits beginning with their 2007 models, and diesel refiners must start selling cleaner fuel by June 2006.

The EPA aims to cut emissions of smog-causing nitrogen oxides and airborne particulate matter. The latter has been tied — inconclusively, say the agency's critics — to health effects ranging from asthma to premature death (see *Nature* 392, 642; 1998). The EPA says that trucks and buses account for about one-third of nitrogen oxides emissions and one-quarter of airborne particulates, or soot.

The American Lung Association welcomed the decision and released poll results suggesting that Americans strongly favour cleaning up diesel exhaust, even if it costs consumers more. The American Petroleum Institute, the US Chamber of Commerce, and several organizations representing farming, oil production, trucking and vehicle manufacturing condemned the new rule,



Diesel vehicles face strict new rules on emissions.

citing the costs of compliance and what they call dubious benefits to public health.

But Richard Kassel, a senior attorney with the Natural Resources Defense Council, an environmental lobby group, says that "the oil

industry is not speaking with a unified voice" on the issue. Indeed, BP and Tosco, two of the nation's biggest diesel fuel producers, support the rule.

During hearings on diesel regulation in September, Senator James Inhofe, a Republican from the oil-producing state of Oklahoma and outgoing chairman of the Senate subcommittee that oversees clean air legislation, threatened to use Congress's powers to veto any regulation he saw as too stringent.

But, says Kassel, the 1996 Congressional Review Act has never been used successfully to overturn a regulation. Inhofe will not head the clean air subcommittee when Congress reconvenes this month. And Republicans may not want to risk reversing the rule.

Environmentalists may also find an ally in Whitman, whose environmental commissioner from New Jersey testified in favour of the rule during hearings.

A group led by the American Trucking Associations has challenged the EPA's 1999 standards on ozone and particulates. So far the courts have ruled against the EPA, and the matter is now being considered by the Supreme Court. ■

Human Proteome Index launched

Alison Abbott

Large Scale Proteomics, based in Rockville, Maryland, this week launched the Human Proteome Index, a database which subscribers can use to help identify proteins involved in diseases.

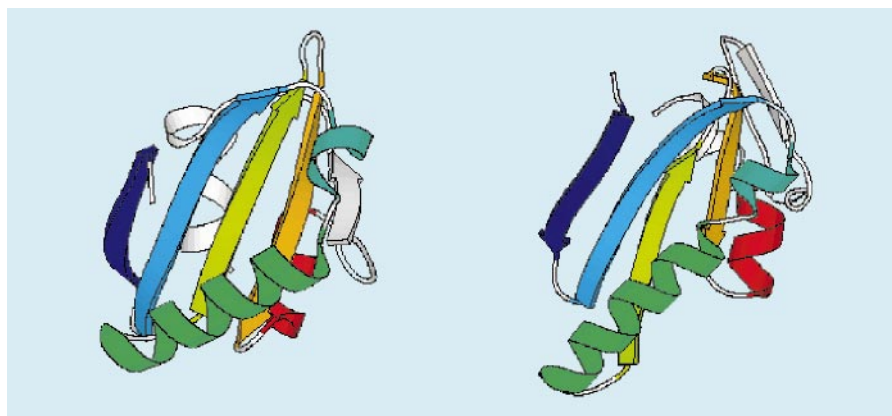
The index — the first of its kind — lists and quantifies the complement of proteins expressed in 157 different tissues from a single female donor who died of cardiac arrest. The proteins were identified 'factory-style' by mass spectrometry after being separated in the different tissues by two-dimensional gel electrophoresis.

The index includes over 115,000 different proteins, encoded by an estimated 18,000 different genes. Humans are generally thought to have between 30,000 and 70,000

genes. Large Scale Proteomics — part of the Large Scale Biology Corporation — is expanding the database, using tissues from additional donors, to detect proteins expressed at low levels and to map the cellular and subcellular distribution of proteins.

A Human Proteome Index was conceived in 1980 by Norman Anderson, then at the Oak Ridge National Laboratory in Tennessee, as a large-scale federally funded project, but the idea lost political support. His son, Leigh Anderson, has been the driving force behind the latest effort.

The Human Proteome Index will only be available commercially, says Leigh Anderson, mainly to companies interested in finding markers of tissue damage or disease. ■



Virtually there: 'native' protein structures (left), are being modelled more accurately by computer (right).

Computer modellers seek out 'Ten Most Wanted' proteins

Alison Abbott

Emboldened by their growing success at using computer models to predict protein structures, computational structural biologists have put out a community-wide call to find the "Ten Most Wanted" proteins whose structures they will try to solve.

The plan to offer modelling services to the wider biological community was hatched at the fourth biannual meeting of the Critical Assessment of Techniques for Protein Structure Prediction (CASP4), held in Asilomar, California, last month.

The computational biologists searching for the top ten believe that computer modelling may become a viable alternative to the current practice of using X-ray crystallography and nuclear magnetic resonance techniques to 'solve' the structure of a protein.

Any biologist can suggest a candidate protein for the Ten Most Wanted list, provided its structure is currently unknown. A website has been created at which candidate proteins can be registered, and where the case for their biological importance can be made.

The ten proteins considered to be the most significant will be modelled. "A protein might make our list if, for example, it is believed to play a key role in cell signalling," says one of the organizers, structural biologist David Baker of the University of Washington.

Protein modellers traditionally pit their prediction skills against each other at CASP meetings. In the year up to a meeting, participants are challenged to model various categories of proteins, and their structures are found experimentally before the meeting.

Although protein-structure prediction is not yet accurate enough to help in drug design, "it has matured to a point where models produced by prediction algorithms can be used to understand and test hypotheses about biological function", Baker says.

Marked improvements were seen in cases

where an amino-acid sequence of interest did not correspond to any known structure. "CASP3 competitors were able to describe 40–60 [amino-acid] residues approximately correctly, whereas at CASP4 even whole small proteins were roughly correct," says one of the meeting's organizers, John Moult, of the Center for Advanced Research in Biotechnology at Rockville, Maryland.

Fully automated modelling, where protein structures are predicted by computer programs without any human input, was also better at CASP4. Predicted structures determined by combining all fully automated approaches were among the top six in the meeting's "fold recognition" category. There are believed to be about 13,000 different ways a protein can fold, and most structural biologists believe that, once the structures of all the folds are understood, models will be able to predict the structure of every protein.

Proponents of fully automated modelling, such as Dani Fischer of Ben-Gurion University, Israel, believe that the pace of improvement is such that humans will soon be beaten by computers "in the way that Big Blue beat chess world champion Kasparov".

Others are more cautious. Moult notes that the importance of automated modelling is "that it is scalable, which is important if you want to go for thousands of proteins", a prospect genomics makes possible. But he says automatic methods are not yet as effective as a combination of human and machine.

The big question, researchers say, is whether modelling will continue to improve fast enough to make an impact. There is a relatively short window of time for structure prediction to be useful before structural genomics — factory-style protein-structure generation using genomic information as starting material — generates all the results needed.

♦ <http://www.doe-mbi.ucla.edu/TMW>

Major UK cancer research charities to consider merger

Georgina Kenyon, London

British cancer researchers could find themselves with a major new research organization, loosely modelled on the US National Cancer Institute (NCI), after the country's two main cancer charities meet this month to consider a merger.

The trustees from the Cancer Research Campaign (CRC) and the Imperial Cancer Research Fund (ICRF) will meet up on 16 January. Even if a full merger is not deemed financially viable, the charities say they will collaborate much more closely in future on basic and clinical research.

"Collaboration is the best way to exploit our resources and reduce our costs," says Gordon McVie, director general of the CRC. Both charities want to make more use of new equipment and techniques, such as microarray DNA chips and X-ray crystallography, he adds.

The CRC currently concentrates on extramural research, whereas most of the ICRF's work is done intramurally at its main institute in London. Paul Nurse, director general of the ICRF, says that a merger would see both classes of work continue in parallel, as happens at the publicly funded NCI in the United States.

But the UK charities have a long way to go in matching the NCI's financial might: each spends just over £100 million (US\$150 million) per year, against the NCI's \$4 billion. "We are chronically short of funds," says McVie. "The British government has abdicated its responsibility to the charities."

Scientists cautiously welcomed discussion of the merger, which could take at least a year to implement. Peter Hoskin, a CRC-funded oncologist at the Mount Vernon Hospital in Middlesex, says: "It would be better if funding and research were less fragmented."



Merger pending: the Imperial Cancer Research Fund's London staff could gain closer ties with university researchers.

ICRF

Study focuses on genetic fallout of the bomb

David Cyranoski, Tokyo

Children born to survivors of the 1945 atomic blasts in Japan are to be the subjects of a new study investigating the long-term effects of the atomic bomb. The research, which gets under way this month, will track the onset of diseases such as heart disease and diabetes, which are unlikely to show up until adulthood.

The work will be done by the Radiation Effects Research Foundation (RERF), a non-profit private foundation funded jointly by the United States and Japan. RERF has laboratories in Hiroshima and Nagasaki — the two cities bombed with nuclear devices in 1945 by the United States.

To date, RERF has found no long-term genetic effects in the so-called F_1 population, the roughly 80,000 children born to survivors of the US attacks between 1946 and 1984. But this result has surprised some researchers, as radiation has shown powerful genetic effects in animal experiments.

One possible explanation could be the technical limitations of the studies themselves, says Seymour Abrahamson, a fruitfly geneticist from the University of Wisconsin-Madison who recently joined RERF as its research director. "We know a lot more now about genetic diseases," he says.

The only systematic clinical study on F_1 children was carried out between 1948 and



Survivors of Hiroshima: tracking their descendants will indicate genetic effects of nuclear war.

1953. Most of these were based on observations by doctors and midwives, who reported stillbirths and physically apparent abnormalities in children less than one year old.

Abrahamson says that there is still a possibility that radiation effects will show up as "late onset, multifactorial" genetic disorders not visible at infancy. "We have to close the circle," he says.

In January, RERF will conduct a postal survey of 20,000 people in the F_1 population to find out how many will be willing to undergo a full-scale clinical analysis. Data on health records and physical examinations will then be collected. It will probably take

between four and six years for results from the analysis to appear, Abrahamson says.

But some of the study's potential subjects are opposed to the research because they fear that finding a genetic effect could lead to discrimination. Others, meanwhile, hope that a link between radiation and ill-health could justify their demands for benefits from the government. Only this year, Japanese courts awarded compensation from the government to plaintiffs who claimed that their illnesses were caused by the 1945 bombing. "This study is likely to influence such court decisions in the future," Abrahamson says.

► <http://www.rerf.or.jp/eigo/experhp/rerfhome.htm>

Parliament gives green light to stem-cell research

David Dickson, London

British scientists reacted with surprise and relief to a substantial vote of confidence from the House of Commons — the lower house of the UK parliament — over the use of human embryos for research into therapies for serious diseases.

Members of Parliament (MPs) agreed by 366 to 174 on 19 December to approve changes to the Human Fertilization and Embryology Act of 1990 allowing such research to take place.

"I was very surprised; we were expecting it to be much closer," says Austin Smith, director of the Centre for Genome Research at the University of Edinburgh, who already holds a licence for carrying out research on human embryos up to 14 days old as permitted by the 1990 act — but is currently restricted to work that is directed at problems of infertility.

Smith, one of the leading UK investigators in stem-cell research, says that a change in the regulations would not have any short-term impact on his research, apart from enabling him to obtain a broader

licence. But it would make it possible for him to start work towards potential therapies, perhaps in neurodegenerative diseases such as Parkinson's and Alzheimer's disease.

The proposed change in the legislation had been strongly opposed both by anti-abortion groups and prominent Roman Catholic leaders. Other opponents warned that the interest on obtaining stem cells from human embryos represents the beginning of a "slippery slope" towards human cloning.

Such arguments had been countered by a vigorous lobbying campaign by the scientific and medical establishment, with MPs being given briefings by, among others, the Royal Society, the Wellcome Trust and the Nuffield Council on Bioethics. The objections were also rejected last summer by a panel set up under the government's chief medical officer, Liam Donaldson (see *Nature* 406, 815; 2000).

But the most telling contributions to the debate perhaps came from several MPs who told of their personal battles with illness,



Supporting stem cells: Aberdeen MP Anne Begg told of her battle against degenerative disease.

including Anne Begg (Labour, Aberdeen South) who suffers from a degenerative brittle-bone disease and is confined to a wheelchair.

The proposed change will next be debated in the House of Lords on 16 January, but, given the size of the Commons majority, researchers are optimistic that the Lords might not reject the measure.

Diet studies send people rushing for comfort food

Washington Many Americans, confronted with a profusion of conflicting studies on diet, have joined a 'nutrition backlash' and embraced unhealthy eating, says a study in this month's *Journal of the American Dietetic Association*.

The study, by researchers at the Fred Hutchinson Cancer Research Center in Seattle, found that over 40% of respondents to a survey were tired of hearing about what foods they should or should not eat. The diet of this group — which included an even higher proportion of the elderly, the poor and young men — contains more fat and less fruit and vegetables than average.

"The majority of the public still care about their diet and health, but there are definitely some sub-groups that have just plain had it," says Ruth Patterson, lead author of the study.

American Physical Society sets up electronic archive

Washington The American Physical Society will this month begin sending a complete electronic archive of eight of its journals to the US Library of Congress for permanent safekeeping.

The archive, which will be updated continually, contains more than a century of physics, dating to the first issue of *Physical Review* in 1893. The society is in the process of electronically scanning older issues of the journal, and members can subscribe to the archive for an annual \$100 fee.

Onsite library users will have free electronic access to all eight archives. In return, the society will get a second repository for its collection.

Australian TV axes science programmes

Sydney Australian scientists are in uproar over the planned closure of *Quantum*, a science show that has run for 16 years on national television, and the axing of the Science TV Unit at the Australian Broadcasting Corporation (ABC).

The Australian Academy of Science, which persuaded the ABC to begin science programmes in 1964, branded the removal of in-house competence as "a leap backwards". Other universities and scientific groups have also attacked the changes.

In a letter to Jonathan Shier, the ABC's new managing director, Peter Doherty, the Australian Nobel laureate in medicine, wrote: "It will be a national tragedy if the capacity of the ABC to offer high quality,



Up in arms: ABC TV and radio staff protest over the axing of science and other programme units.

in-depth reporting of the arts and sciences is further compromised." But Shier says that "drop dead day", his own description of the proposed restructuring, will enhance science coverage on television.

Bioethics body calls for reform of clinical trials

Washington The US National Bioethics Advisory Commission (NBAC) has advised the federal government to spend more on broader oversight to protect research subjects.

The NBAC's report calls for Congress to create a single agency that would monitor all clinical trials. The committee also urged federal sponsors of biomedical research, such as the National Institutes of Health, to spend more on monitoring clinical trials.

The report was released at the end of last month and is available for public comment until 17 February.

► <http://www.bioethics.gov>

Hungary gives science budget a 50% boost

Budapest The Hungarian government last month approved an increase in public spending on science and technology of more than 50%. The move reinforces Hungary's efforts to become a leader in science in central and eastern Europe.

Over the next two years, a rise in the science budget of 140 million euros (\$US132 million) will fund higher salaries and more research grants. Private companies will be allowed to write off double what they spend on research and development against tax.

Norbert Kroo, secretary-general of the Hungarian Academy of Sciences, says that Hungary's participation in the European Union's Framework Programme "paved the way" for the research boost.

Awards for Mediterranean geneticist and physicist

Paris André Megarbane, a geneticist at the Université Saint Joseph in Beirut, and Abderrahmane Tadjeddine, deputy director of the LURE synchrotron laboratory in Orsay, France, have jointly won the 2000 Rammal Award for their contributions to scientific collaboration in the Mediterranean region.

The award, named after the Lebanese physicist Rammal Rammal, is made each year by Euroscience, an association that promotes European science.

Megarbane was selected for his research in the Lebanon and his collaborations with scientists in ten other Mediterranean countries. Tadjeddine was cited for supporting education in Algeria and his role in plans to relocate a synchrotron from Germany to Jordan.

British BSE committee opens up to the public

London Advisors to the UK government are planning to open up their discussions on mad cow disease and variant Creutzfeldt-Jakob disease (vCJD) to the public for the first time.

A recent public enquiry into the BSE epidemic found that a culture of secrecy had contributed to its spread (see *Nature* 408, 3–5; 2000). To help regain public confidence, some members of the Spongiform Encephalopathy Advisory Committee — a group of experts who advise ministers on the science of BSE and vCJD — are pushing to have some of their sessions made public. Sessions involving commercial or patient confidentiality will remain closed.

Currently, press briefings are held after each meeting of the committee. But Peter Smith of the London School of Hygiene and Tropical Medicine, the committee's deputy chairman, says that it will meet in public during 2001.



Going public: the inquiry into the British BSE epidemic condemned a 'culture of secrecy'.

Mind

the pseudogap

Physicists are still searching for a convincing theory of high-temperature superconductivity. But at least the nature of the puzzle is becoming clearer. Mark Buchanan weighs the odds of a breakthrough in understanding.

We can blame Georg Bednorz and Alex Müller. Fifteen years ago, physicists thought they understood superconductivity — the tendency of some materials to conduct current with zero resistance at temperatures of around 10 to 20 kelvin. But in 1986, working at the IBM Research Laboratory in Zurich, Bednorz and Müller unveiled a new class of ceramic materials that pulled off the trick at much higher temperatures¹. More than 100,000 research papers later, and with the record for high-temperature superconductivity now standing at some 150 K, theorists are still baffled.

"We now probably know more experimentally about this class of materials than any other," says Peter Littlewood, a condensed-matter physicist at the University of Cambridge. "But we still have a fundamental conundrum."

In the past few years, however, as physicists have found ways to follow what electrons are doing inside high-temperature superconductors with an accuracy once thought

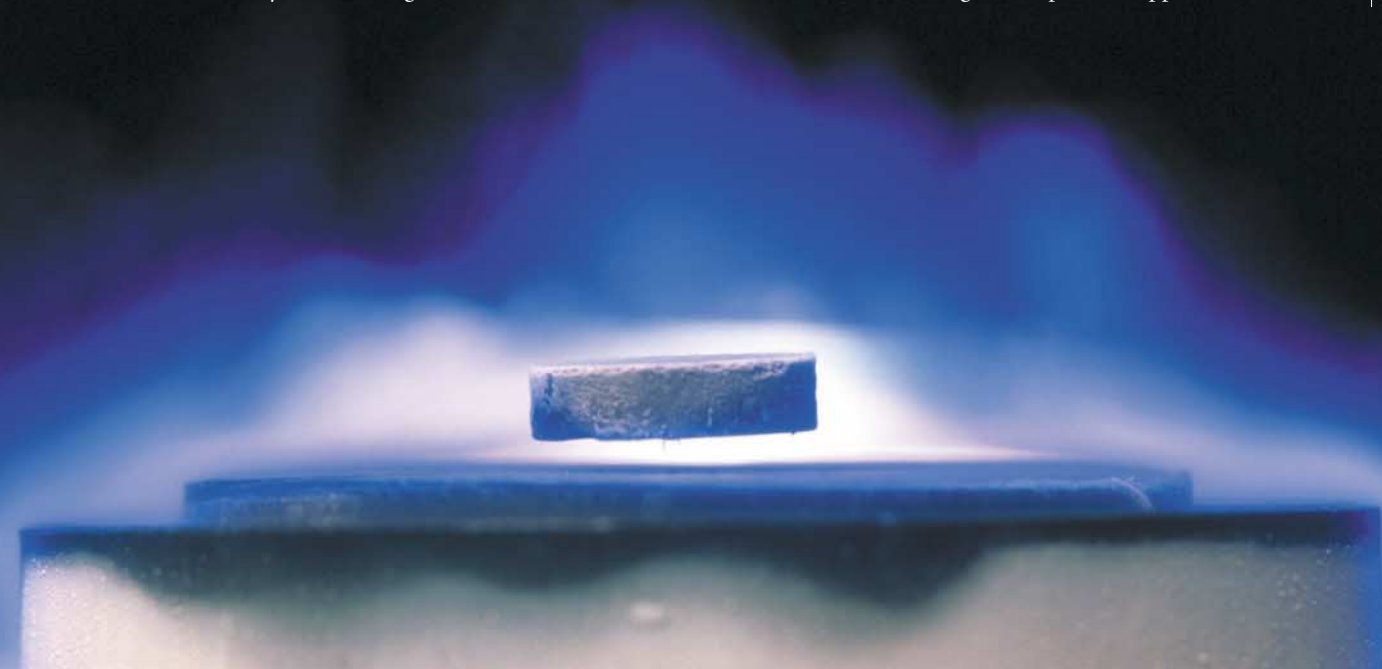
impossible, there is growing optimism that this conundrum will be resolved. "There will be a glorious theory of physics at the end of the journey," predicts Jan Zaanen, a theoretical physicist at Leiden University in the Netherlands. And the key may be an understanding of how these ceramics behave when heated just beyond the temperatures at which they superconduct — in particular a curious phenomenon known as the 'pseudogap'.

The X-factor

Ordinary superconductors are simple metals. But the high-temperature variety discovered by Bednorz and Müller are more complicated. Known informally as cuprates, they are made up of parallel two-dimensional sheets containing copper and oxygen atoms. Sandwiched between these sheets are other kinds

of atoms — in the material $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$, for example, these atoms are lanthanum and strontium. Within the two-dimensional sheets, each copper atom has the potential to relinquish one loosely bound electron which can then move from one copper atom to another and carry electricity. What actually happens depends on the value of x .

With no strontium ($x=0$), the material is a poor conductor and is known as a 'Mott insulator'. Because electrons repel one another quite strongly, the energy needed to force a second electron onto a copper atom that has not already relinquished its loosely bound electron is prohibitive. So with one loosely bound electron on virtually every copper atom, the resulting electronic traffic jam makes the material a poor conductor. This 'undoped' material is also antiferromagnetic — an ordered state with no bulk magnetization. This occurs because each electron carries a spin and acts as a tiny magnet. The forces between these magnets tend to make neighbours point in opposite directions, so



the spins point alternately up and down across the material.

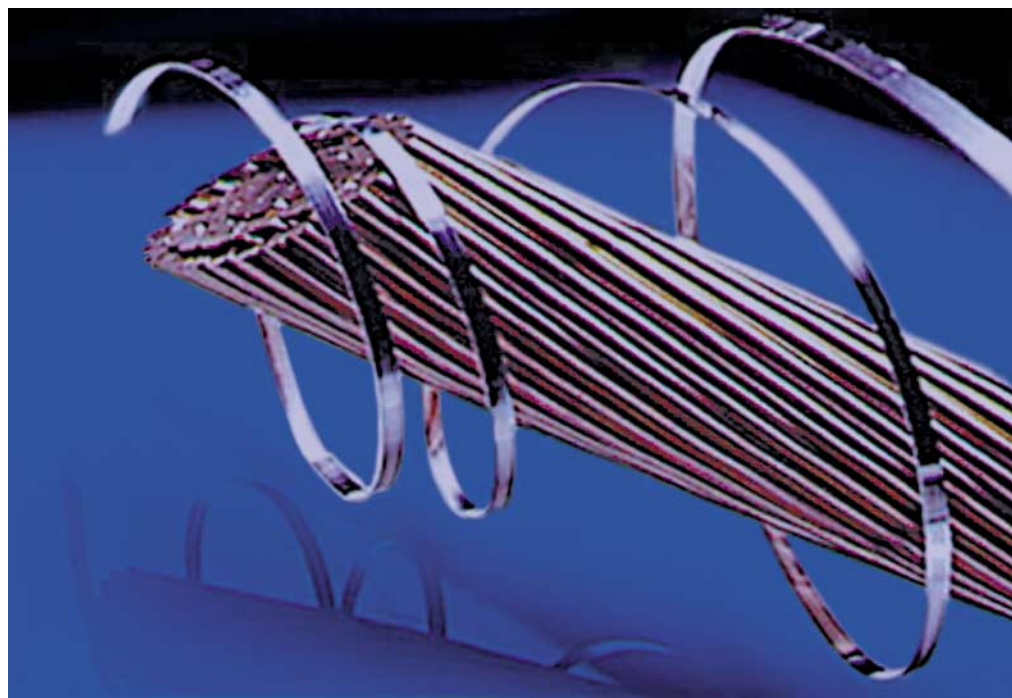
Things get more interesting when strontium atoms replace some of the lanthanum atoms. Strontium has a greater affinity for electrons and so attracts the loosely bound electrons from the copper atoms. With some copper sites now vacant, electrons can hop between the copper atoms and so carry electricity. When the material is doped with strontium, and x increases from zero, $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ first turns from an antiferromagnetic insulator into a reasonable conductor. At larger values of x , the material gains the ability to be superconducting when cooled. The transition temperature, T_c , at which superconductivity arises depends on the value of x , reaching a peak at around $x = 0.15$ (see figure, below).

All the cuprates show this same basic behaviour, and a good theory of high- T_c superconductivity would explain why. No such theory yet exists, but when it does, it is sure to be very different from the theory of conventional superconductors. This was created in the late 1950s by John Bardeen of the University of Illinois, Leon Cooper of Brown University in Rhode Island and J. Robert Schrieffer of the University of Pennsylvania. Known as the BCS theory, it stands firmly on the shoulders of an earlier theory of 'quantum liquids', devised by Lev Landau of the Institute of Physical Problems in Moscow.

Storm clouds

Inside an ordinary metal, a cloud of electrons moves about within a crystal lattice of atoms and ions. These electrons interact strongly with one another, like the particles in a liquid, and the movement of any one will stir up a local storm in the motions of others. This made theoretical analysis seem nearly impossible. But Landau showed that, within the setting of a quantum liquid, the electron plus the storm would act as a 'quasi-particle' in its own right. So the electronic properties of metals can be understood fairly simply in terms of a collection of quasi-particles.

Bardeen, Cooper and Schrieffer used this picture to unravel the mystery of superconductivity. They realized that in some metals the motion of one quasi-particle might set the lattice of atoms and ions vibrating in such a way as to exert an attractive force on other quasi-particles, even binding the quasi-particles together into pairs. This has far-reaching consequences. On their own, the quasi-particles obey the Pauli exclusion principle — no two can occupy the same quantum state. But when paired together they no longer face this prohibition. And, at low enough temperatures, the pairs all condense into the same quantum state. This condensation is the key to superconductivity. Because the condensate of pairs is a single quantum entity, individual pairs are no



Got it taped: superconducting ribbons can carry as much current as copper cables with 100 times the cross-sectional area. Like all superconductors, cuprates can be levitated over a magnet (opposite).

longer scattered by localized impurities in the material. They can therefore glide along without losing energy.

Unfortunately, this nice picture falls to pieces when faced with the cuprates. Here, too, the charge carriers — which might be electrons, or the 'holes' left behind when electrons are displaced — are generally thought to move around in pairs. But there is a big difference. In the BCS theory, when the lattice vibrations force the pairs together, they form without any internal angular momentum — the charge carriers within each pair do not rotate about one another. In the cuprates, there is angular momentum. In the theorists' jargon, this is 'd-wave', as opposed to 's-wave' superconductivity. In addition, lattice vibrations should not be strong enough to keep the

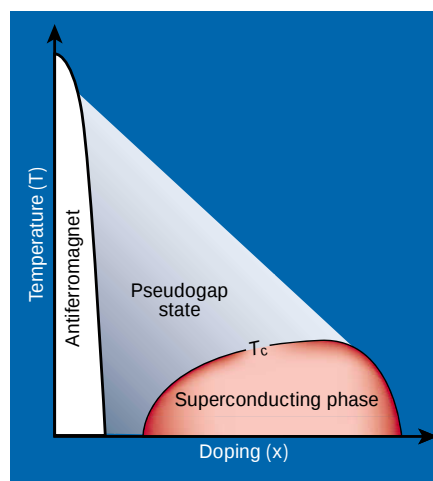
pairs together and give superconductivity much above 30 K — way too low to explain what happens in the cuprates.

Shooting gallery

To probe the properties of cuprates, condensed-matter physicists turned to a technique called angle-resolved photo-emission (ARPES). In this method, researchers shoot high-energy photons at a material, and then measure the energy and momentum of any electrons that get knocked out. These values are nearly identical to that of the stimulating photon, but tiny differences reveal how much energy and momentum the material absorbed. These differences also characterize the material's natural quasi-particles. Given the precise measurements needed, says Andy Millis, a condensed-matter theorist at Rutgers University in New Jersey, "it is amazing that it works". But work it does — and in the past five years or so, the technique has revealed strange behaviour in the cuprates.

When superconducting, any ordinary superconductor has an 'energy gap'. If you fire photons at the material in an effort to excite quasi-particles, nothing much happens until the input of energy rises above a minimum value known as E_g — this being the energy needed to break apart one of the charge-carrier pairs. Not surprisingly, this energy gap goes away when the temperature rises above T_c and the pairs again split into single particles.

The cuprates also have an energy gap, but things work out rather differently. Below T_c , ARPES experiments reveal that the superconducting gap depends on the direction in



Bridge that gap: the temperatures at which a cuprate superconducts are affected by doping.

which the photon is fired into the material. There is a large gap in some directions and a small gap in others^{2,4}. This just reflects the 'd-wave' nature of the pairs and fits in well with Landau's theoretical picture. But things go terribly awry when you heat the materials up above T_c — the energy gap does not go away^{5,6}.

This anomaly is the pseudogap, and it makes a shambles of the conventional BCS theory. Other features of the ARPES results for cuprates are even more disquieting. Below T_c , if you put in enough energy to overcome the energy gap, you can excite quasi-particles. These show up as a sharp peak in the data, as the material absorbs energy in the conceptually simple way that Landau said it should^{5,6}. But above T_c , the sharp peak becomes broad and confused, as if the quasi-particle concept does not apply⁷. "Well-defined quasi-particles have not been observed in the normal state of any of the cuprates," says Zaanen.

This picture has emerged as data from dozens of different high-temperature superconductors have come to tell the same story. In the past, says John Tranquada, an experimentalist at the Brookhaven National Laboratory on Long Island, New York, "it has often been easier to attribute disturbing properties to 'bad samples' than to change one's perspective". But physicists are now facing up to the cuprates' undeniable peculiarities. And with the traditional BCS theory in ruins, they are looking elsewhere for guidance.

Coherent thoughts

The pseudogap may provide one key. In 1995, Vic Emery of the Brookhaven National Laboratory and Steve Kivelson of the University of California at Los Angeles pointed out that superconductivity requires more than just paired charge carriers — it also requires 'phase coherence' between those pairs⁸. Loosely speaking, each pair has a quantum wave associated with it, and for the pairs to condense into the superconducting state all the waves have to be in phase with one another. As the pseudogap exists almost up to room temperature, it could be that some feature of cuprate structure makes it possible for pairs to form at high temperatures, well above T_c . If so, the onset of superconductivity would signify not the formation of pairs, but the setting in of phase coherence below T_c .

In this view, cuprate superconductivity may break down above T_c because the pairs have so much thermal energy that they can no longer maintain phase coherence. Emery and Kivelson suggest that, just above T_c , superconductivity should therefore become fragmented or fluctuating — it should be possible to find evidence of superconductivity in the material, but only over very short distances or timescales.

In 1999, this idea won support from experiments by a team led by Joe Orenstein

of the University of California at Berkeley⁹. Orenstein and his colleagues studied how another type of cuprate — $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ — responded to an electrical field that was alternating very rapidly. This allowed them to test Emery and Kivelson's idea by probing the material's properties over the appropriate distances and timescales. They found exactly the sort of fluctuations that Emery and Kivelson had predicted. Above T_c , the higher levels of thermal energy appeared to churn up small 'vortices' — regions in which the material became non-superconducting. And the number of these vortices grew with temperature, ultimately taking over and destroying superconductivity completely.



In August last year, Shin-ichi Uchida of the University of Tokyo and his colleagues found further evidence for such vortices¹⁰. They placed a thin cuprate film into a magnetic field, and made one edge warmer than the other. Magnetic fields generally will not pass through a superconductor, but they can penetrate through non-superconducting vortices. By looking for evidence of the magnetic field in the material, Uchida and his colleagues showed that vortices formed above T_c . This again suggests that the charge carriers in cuprates do become paired at high temperatures, and then turn superconducting when things get colder.

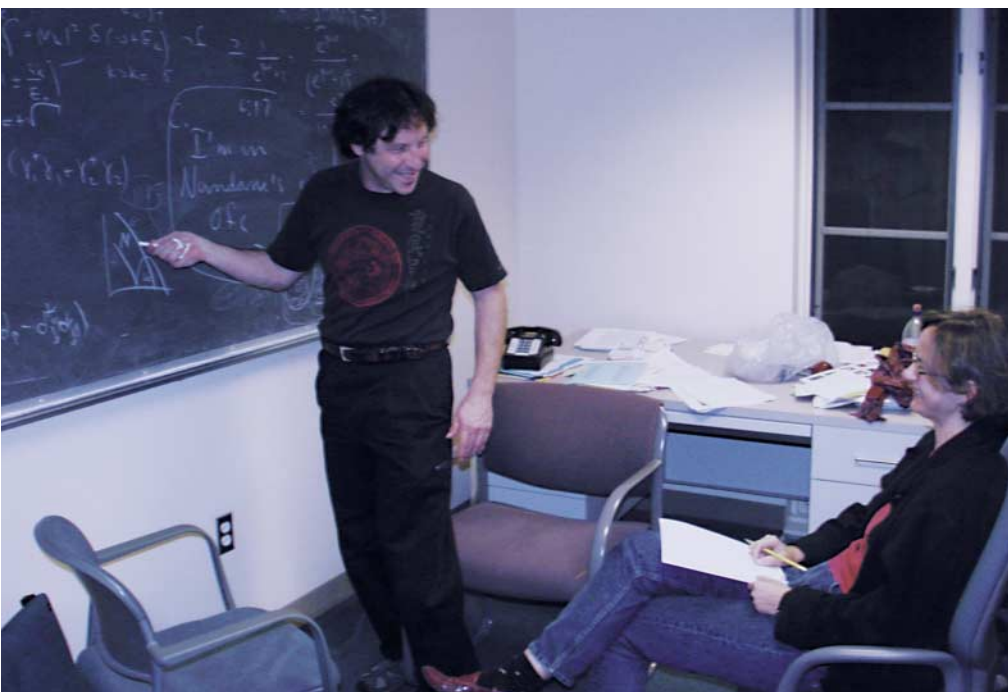
So what might the pairing mechanism be? Most theorists are now focusing on the physics that takes place in a Mott insulator in which just a few electrons have been removed from the copper atoms, creating a small number of charge 'holes'. "Magnetic interactions play an important role in every theory," says Millis. In undoped cuprates, the magnetic interactions between the electrons force an antiferromagnetic arrangement of their spins. These interactions must change when electrons start to move around, with the pairing of charge carriers being one consequence. The puzzle is to find out why this occurs.

Spins and stripes

Some researchers believe that fluctuations in electron spins might play a role similar to the lattice vibrations in the conventional BCS theory — the spin of one moving electron might distort nearby spins in such a way as to exert an attractive force on other electrons, binding them together into pairs^{11,12}. But others suspect that something more exotic may be going on. One popular idea centres on theoretical calculations that suggest that in a lightly doped Mott insula-



Pioneers: Bednorz (top) and Müller received a Nobel prize for their discovery of high-temperature superconductors, materials that have caused many theoretical physicists sleepless nights.



Unfazed: Steve Kivelson sees quantum wave coherence as key to high-temperature superconductivity.

tor — in which a few electrons have been removed from the copper atoms to break the electronic traffic jam — the charges and spins will undergo a natural process of segregation.

Charge holes can move from atom to atom, but have trouble doing so in a largely antiferromagnetic material. “Any movement tends to create strings of misaligned spins,” says Emery. This tends to force the holes into ‘stripes’ — linear regions with lots of holes, like rivers of charge, running parallel to other regions with few holes and electron spins locked into the antiferromagnetic pattern.

Based on this picture, Emery and Kivelson theorize that superconductivity would emerge in several stages¹³. Being forced together into the stripes, the holes can become bound together into pairs even at high temperatures — thus generating the pseudogap. At a somewhat lower temperature, the pairs within each stripe would then condense into single quantum states, giving rise to superconductivity — but only in the stripes. Finally, at T_c and below, the weakening of thermal noise would permit the pairs to hop between stripes, giving phase coherence and full superconductivity.

In neutron-scattering experiments, researchers have now seen stripes of this sort in several different types of cuprates, doped to a variety of levels at a range of different temperatures¹⁴. Yet for the Emery–Kivelson theory to work, when the material is superconducting, the stripes have to vibrate or meander within the two-dimensional copper-oxide planes. Theoretical calculations suggest that stripes should work against superconductivity if they are more or less stationary¹⁵. And although some experiments offer hints that

fluctuating stripes may exist in the cuprates¹⁶, the significance of this for superconductivity remains unclear. “Most people seem to accept that stripes can occur in certain compounds,” says Tranquada. “But there is still considerable scepticism as to whether fluctuating stripes are a universal feature of the cuprates.” Littlewood agrees: “The real question is whether the stripes are central to the problem or instead represent a red herring.”

Fractional progress

One of the most exciting recent alternatives to the stripe picture comes from theorists Matthew Fisher and Todadri Senthil of the University of California at Santa Barbara. Fisher and Senthil suggest that, when you give energy and momentum to an electron, it splinters into constituents which behave as free particles — some, called ‘spinons’, carrying spin and others, called ‘chargons’, carrying charge. This ‘fractionalization’ of the electron would explain why cuprates in their non-superconducting state appear to lack well-defined quasi-particles. This view also predicts that charge and spin should come back together with the onset of superconductivity, and so sharp quasi-particles should then exist again — as they do, at least in the direction where the pseudogap is largest. “No other theory seems able to account for this,” says Millis.

The basic idea of electron fractionalization has been around for years¹⁷. But some physicists believe that Fisher and Senthil have found a more natural theoretical framework in which it might occur^{18,19}. “Many of us see this as a demystification of the notion of spin–charge separation,” says Zaanen.

One of the theory’s most intriguing

aspects is that it predicts superconductivity might even take place without any pairing between charge carriers. Superconductivity requires the charge carriers to condense into a single quantum state — and the Pauli exclusion principle means that pairing is usually the only way to achieve this. But the separation of charge and spin releases the chargons from this restriction, allowing them to condense into a single quantum state without pairing ever taking place.

Fisher and Senthil say that their ideas can be put to the test. “They have proposed a spectacular experiment which can prove or disprove it,” says Zaanen. In addition to the chargons and spinons, the theory proposes that electron fractionalization would cause the formation of a third type of excitation called ‘visons’ — vortex-like anomalies that can associate with magnetic fields. If visons exist, it should be possible to trap them by heating and cooling a cylindrical cuprate sample with a hole drilled down the centre, and turning a magnetic field on and off, all in a precise sequence. The visons’ presence would be betrayed by a small portion of magnetic field becoming trapped in the central hole.

Although both stripes and electron fractionalization have appeal, no one yet knows whether the right ideas are yet on the table. What is clear is that the high-temperature superconductors live well outside the regime of traditional theories. And it could be that there will never be a simple way to understand them. “We might ultimately find a theory that everyone believes,” says Littlewood, “even though no one will be able to use it to calculate anything.”

But after all this effort, most theorists would prefer to think that someone, someday, will come up with a new perspective that will suddenly make everything fall into place. “I wait with bated breath for some young kid to come up with a completely new idea,” says Littlewood. ■

Mark Buchanan is a science writer in Normandy, France.

1. Bednorz, J. G. & Müller, K. A. *Z. Phys. B* **64**, 189–193 (1986).
2. Shen, Z. X. *et al. Phys. Rev. Lett.* **70**, 1553–1556 (1993).
3. Kelley, R. J., Ma, T., Quitman, C., Magarito, G. & Onellion, M. *Phys. Rev. B* **50**, 590–593 (1994).
4. Ding, H. *et al. Phys. Rev. Lett.* **74**, 2784–2787 (1995).
5. Loeser, A. G. *et al. Phys. Rev. B* **56**, 14185–14189 (1997).
6. Federov, A. V. *et al. Phys. Rev. Lett.* **82**, 2179–2182 (1999).
7. Orenstein, J. & Millis, A. J. *Science* **288**, 468–474 (2000).
8. Emery, V. J. & Kivelson, S. A. *Nature* **374**, 434–437 (1995).
9. Corson, J. *et al. Nature* **398**, 221–223 (1999).
10. Xu, Z. A., Ong, N. P., Wang, Y., Kakeshita, T. & Uchida, S. *Nature* **406**, 486–488 (2000).
11. Scalapino, D. J. & White, S. R. *Phys. Rev. B* **58**, 8222–8224 (1998).
12. Scalapino, D. J. *Science* **284**, 1282–1283 (1999).
13. Emery, V. J., Kivelson, S. A. & Tranquada, J. M. *Proc. Natl. Acad. Sci. USA* **96**, 8814–8817 (1999).
14. Tranquada, J. M., Mallozzi, R., Orenstein, J., Eckstein, T. N. & Bozovic, I. *Nature* **375**, 561–563 (1995).
15. Kivelson, S. A., Fradkin, E. & Emery, V. J. *Nature* **393**, 550–553 (1998).
16. Mook, H. A., Dai, P., Doğan, F. & Hunt, R. D. *Nature* **404**, 729–731 (2000).
17. Anderson, P. W. *Science* **235**, 1196–1198 (1987).
18. Senthil, T. & Fisher, M. P. A. *Phys. Rev. B* **62**, 7850–7881 (2000).
19. Senthil, T. & Fisher, M. P. A. <http://xxx.lanl.gov/abs/cond-mat/0008082>

Global biodiversity plan needs to convince local policy-makers

Sir — The symposium of eminent ecologists¹ that backed the Conservation International (CI) blueprint to save global biodiversity (covering protection of biodiversity hotspots² and tropical wilderness areas) concluded that ecological criteria can be used to set global conservation priorities, and that “all conservation must be driven by what a country’s people want, not by what developed nations impose from outside”³. The degree to which development priorities of target countries are consistent with the CI blueprint will therefore determine its efficacy.

Two of the 25 global hotspots and one of CI’s three focal wilderness areas (New Guinea) cover Indonesian territory. During 1999, I surveyed 125 professionals living in Indonesia and active in development of national biodiversity policy. Based on pilot surveys, I formulated seven questionnaire statements on biodiversity policy and asked respondents to mark the two issues that concerned them most and the two that concerned them least. The response rate was 84% (Indonesian, 74; expatriates, 31). Full details of the survey can be found at <http://www.geog.ox.ac.uk/research/bie/papers/indosurvey.pdf>.

In descending order of concern, the score for each statement is as follows: (1) creating responsible use of land and renewable natural resources to provide quality and sustainable livelihoods (most concern 45; least concern 14); (2) developing informed and rational planning and decision-making (39; 17); (3) maintaining the ecological processes, services and benefits that underpin human activity (31; 10); (4) empowering local communities in natural resource management (30; 34); (5) conserving species and habitat diversity (20; 18); (6) ensuring that weaker communities do not suffer the effects of environmental degradation created by more affluent sectors of society (17; 50); and (7) preserving native forests and other wild places (11; 32).

The CI blueprint closely reflects the issue statements ranking fifth (biodiversity hotspots) and last (wilderness areas). This suggests that CI will need to invest substantially in efforts to convince biodiversity policy-makers in Indonesia that they want its blueprint.

The management of natural resources is likely to be a key issue in the public

debate on the future shape of Indonesia as a nation state, on account of the widely held view that the autocratic Suharto regime (1966–98) plundered natural resources.

At first sight, a promising approach for CI might be to advocate synergistic linkages between its own agenda and the two highest-ranking concerns in my survey — sustainable development and rational decision-making. This is the vision of the critical ecosystem partnership fund recently launched by CI together with the World Bank and the Global Environment Facility⁴.

However, just such an approach has been pursued in Indonesia since the early 1980s, with little gain for biodiversity conservation⁵. This is because major threats to biodiversity derive from large public and private investments, not the local communities who are the focus of concern of Indonesian policy-makers and international donor assistance.

A second obstacle to the CI blueprint is the fact that ‘wilderness’ protection is largely a North American concept, alien to Indonesians.

Ecologists do not doubt the urgency and importance of conserving species and tropical wilderness areas. CI’s initiative is therefore welcome. However, translating the blueprint into action will require that policy-makers in targeted regions give CI’s ecological goals higher priority than is currently the case in Indonesia, where scientific and economic arguments for species conservation have had limited impact on the ground.

Paul Jepson

*School of Geography and the Environment,
University of Oxford, Mansfield Road,
Oxford OX1 3TB, UK*

1. Dalton, R. *Nature* **406**, 926 (2000).
2. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. *Nature* **403**, 853–858 (2000).
3. www.defyingnature.org (accessed 9 November 2000).
4. www.cepf.net (accessed 23 October 2000).
5. Wells, M. et al. *Investing in Biodiversity* (World Bank, Washington DC, 1999).

Is this the first portrayal of tool use by a chimp?

Sir — We recently took part in the Royal Society’s New Frontiers in Science 2000 Exhibition, presenting an exhibit based upon our report in *Nature* of extensive cultural variation in tool use and other behaviour in wild chimpanzees¹. The primary aim of this annual exhibition is to communicate the work of scientists to the public, but we found information flowed in both directions.

Of particular note, following the appearance of the exhibition at the Royal



Society of Edinburgh, was that Peter Sharp drew our attention to a Liberian postage stamp that was issued in 1906 (see above).

Although primatology appears to have been unaware of it, the image may be the earliest accurate depiction of tool use in chimpanzees (perhaps in any animal), antedating by at least half a century the photographic records that later illustrated the first systematic scientific accounts of such behaviour, published in *Nature* by Jane Goodall².

By contrast with earlier depictions, the image is remarkably correct in overall anatomy and posture, including knuckle-walking with the left hand and plantigrade positioning of the feet. Approaching a termite mound (*Macrotermite*), the chimpanzee wields a stout stick of the dimensions later documented as associated with digging into such mounds to access the termites^{3,4}. We know of no scientific report of termite-digging for Liberia.

The basis on which the image was composed remains unknown.

Michael Harvey of the philatelists Stanley Gibbons Limited has identified the stamp as one of a set printed in London by Perkins Bacon, a now defunct company. The stamp appeared as an illustration in *Liberia* (1906), by the naturalist and philatelic illustrator, Sir Harry Johnson, but no clues to the origin of the image appear there, nor in two earlier volumes he drew on heavily^{5,6}. Thus, the trail has gone cold.

Any assistance from your readers in tracing the origins of this scientifically intriguing image will be gratefully received.

Andrew Whiten*, William C. McGrew†

**Scottish Primate Research Group, School of Psychology, University of St Andrews, St Andrews KY16 9JU, Scotland*

†Departments of Anthropology and Zoology, Miami University, Oxford, Ohio 45056, USA

1. Whiten, A. et al. *Nature* **399**, 682–685 (1999).
2. Goodall, J. *Nature* **201**, 1264–1266 (1964).
3. Jones, C. & Sabater-Pi, J. *Nature* **223**, 100–101 (1969).
4. McGrew, W. C. *Chimpanzee Material Culture: Implications for Human Evolution* (Cambridge Univ. Press, Cambridge, 1992).
5. Büttikofer, J. *Reisebilder aus Liberia* (E. J. Brill, Leyden, 1890).
6. Dapper, O. *Description de l'Afrique* (Chez Wolfgang, Waesberge, Boom & van Someren, Amsterdam, 1686).

1901 and all that

The last will and testament of Alfred Nobel provides rich pickings.



known to its old inhabitants. Our prize goes to another northern novelty, the prize of science prizes, first presented 100 years ago this coming 10 December, after the foundation established to effect the will of Alfred Nobel had settled with his relatives and girlfriend.

We begin our list of events recommended for anniversarial recognition this year with centenaries and move backwards, at 50-year intervals, to our most remote occurrence in 901. We then offer Nobel half- and quarter-centenaries taken from the defunct 20th century. As usual, '¢' indicates a period of 100 years and '\$' indicates a Nobel prize.

1901 (1.0 ¢)

Three of the most eminent and influential leaders of the Century of Physics — Enrico Fermi, Werner Heisenberg and Ernest Lawrence — all first saw daylight in 1901. Heisenberg invented quantum mechanics in its matrix formulation, for which he received the Nobel prize (\$1932); Fermi created a cornucopia of new isotopes by irradiating all the elements he could find with slow neutrons, for which \$1938; Lawrence invented the cyclotron and let others use it to make artificial radioactive elements, whence \$1939.

Although Heisenberg, Fermi and Lawrence all did their prizewinning work at universities (Göttingen, Rome and Berkeley, respectively), each played an important part in the advisory apparatus of government. That was a consequence of their performance in wartime projects to release nuclear energy. Lawrence was one of the chief engineers of the Manhattan Project. Fermi demonstrated the feasibility of producing plutonium on an industrial scale. Heisen-



Hats off: John D. Rockefeller, left, established his eponymous biomedical institute a century ago.

berg directed the German uranium project to a purpose still clouded by the uncertainty principle, or lack of principle, that then governed his conduct.

Leading physicists also died in 1901. George Francis FitzGerald (born 1851, 1.5 ¢), proposer of the Lorentz–FitzGerald contraction; Zénobe Théophile Gramme, inventor of the ring dynamo (1871, 1.3 ¢); and Henry Augustus Rowland, deviser of the optical grating and fierce defender of pure physics in practical-minded America.

The interdisciplinary grim reaper also made possible centennial observances of Charles Hermite, the prolific French mathematician; François-Marie Raoult, the French chemist whose determination of specific heats became an exemplar of precise measurement; Max von Pettenkofer, the Munich hygienist who had earlier escaped after drinking a cocktail of cholera bacilli to show that they alone did not cause the disease; Giulio Cesare Bizzozzo, the Italian microscopist who described red blood cells forming in the bone marrow and platelets circulating in the bloodstream; and Joseph Le Conte, the American naturalist who tried to make evolution palatable to clergymen.

One clergyman — and a Baptist one at that — who developed a taste for biomedical

J. L. Heilbron and W. F. Bynum

Doomsayers and premature celebrants can now unite with sober calendrists in agreement that whenever the new millennium began, it is here. That recommends an upbeat Anniversary of the Year, hence not the death of the hard-nosed Dane Tycho Brahe (1601), despite his provision of the data on which modern astronomy was to be built, nor that of his countryman Hans Christian Oersted (1851), whose discovery of the magnetism of an electric current made a great step towards the unification of physical theory and the invention of our technological society.

We choose rather birth: not that of the Swede Anders Celsius (1701), although his scale of temperature regulates our lives, nor that of Christopher Columbus (1451), who — continuing this Scandinavian theme — sailed in the wake of one (Leif Erikson, *circa* 1001) to discover a new world already well



Bomb squad: Lawrence (right, with J. Robert Oppenheimer) would work on the Manhattan Project.



Horses were bled to obtain diphtheria antitoxin.

research, if not evolution, was Frederick T. Gates. As John D. Rockefeller's principal adviser on philanthropy, Gates undertook to convince Rockefeller that if doctors knew more, they could do more. Although Rockefeller himself frequented homoeopaths, he listened. One consequence was the Rockefeller Institute for Medical Research (now Rockefeller University), a nursery for Nobelists established a century ago.

In the United Kingdom alone, in 1901, 630,000 people died and 1,060,000 were born. To help in diminishing the excess, Friedrich Alfred Krupp of Essen invented a fast-acting cannon, and Kristian Birkeland of Oslo an electrical one. As a help to understanding the excess, Havelock Ellis published the first volume of his *Studies in the Psychology of Sex*, completed in six volumes in 1910, and Sigmund Freud exposed other distractions to which flesh is heir in his *Psychopathologie des Alltagslebens* (*Psychopathology of Everyday Life*), which gave the world the concept of the Freudian slip. To help ensure that the newborns were of the right sort, Karl Pearson established his eugenical journal, *Biometrika*.

Scientists a century ago spent much of their time with radiation. In 1901 Guglielmo Marconi (\$1909) transmitted the letter 'S' from Cornwall to Newfoundland; Carl Ferdinand Braun, who shared the prize with Marconi, deployed a crystal to detect radio waves; Pyotr Nicolayevich Lebedev measured the pressure of light; and Julius Elster and Hans Geitel recorded rays from radioactive material in the soil and atmosphere. Meanwhile, Willem Einthoven (\$1924) began recording another kind of wave, those made by the contracting heart, on his string galvanometer.

Nobel would have been pleased with the honouring of Marconi and Braun. Their contributions to wireless telegraphy were the sort of thing he had in mind in directing that his prizes should go to people who "during the preceding year shall have conferred the greatest benefit on mankind". The first winner in physics, Wilhelm Conrad Röntgen of Würzburg, also met Nobel's terms perfectly. His X-rays, although replete with long-term problems in physics, found immediate and vital employment in medicine. The other prizewinners for 1901 were, for chemistry,

Jacobus Henricus van't Hoff (died 1911, 0.9 ¢) of Berlin and, for medicine, Emil von Behring of Marburg. Van't Hoff's benefits to mankind were the laws of osmotic pressure and chemical dynamics; von Behring's, the development of serum therapy, which he mixed with a thoroughly modern desire to exploit his discoveries commercially.

1851 (1.5 ¢)

Fears that it would be a playground for all the pickpockets, prostitutes and vagabonds of Europe proved unfounded — the Great Exhibition held in London's Hyde Park in 1851 drew so many respectable people that it made money. The proceeds helped endow the Museum of Natural History, the Science Museum, the Victoria and Albert Museum, the Royal College of Chemistry, and a number of 'exhibitions': scholarships for the study of science. Among many distinguished exhibitors was Ernest Rutherford (born 1871, 1.3 ¢, \$1908), whom it enabled to study in Cambridge. The immediate high point of the exhibition was the glass and iron building designed by Joseph Paxton in which it took place. This Crystal Palace housed, along with trees growing in the park, displays of the innovating technology of the day. Yankee ingenuity, as incorporated in the machines designed by Cyrus McCormick and Obed Hussey for reaping, and by Isaac Singer for sewing, made the greatest impression; but the shape of things to come was indicated better by the display of devices associated with the electrical telegraph, such as electric fire alarms and armoured submarine cable.

The induction coil invented in Paris in 1851 by Heinrich Daniel Ruhmkorff became the supplier for experiments with gaseous discharges, from which the discoveries of X-rays and the electron were to come. Among the perfectors of the Ruhmkorff coil was Armand-Hippolyte-Louis Fizeau, who began his career of perfection by improving the photographic process of Louis-Jacques-Mandé Daguerre (died 1851, 1.5 ¢). Fizeau's detection of the dragging of light by moving water also dates from 1851, as does the pendulum for demonstrating the Earth's rotation devised by his friend Jean-Bernard-Léon Foucault.

It was an eventful year for the eye. The physician-turned-physicist Hermann von Helmholtz invented the ophthalmoscope. Even without Helmholtz's instrument, Heinrich Müller discovered visual purple (rhodopsin) and Ferdinand von Arlt described the trachoma that bears his name, as well as an operation to correct distichiasis, an anomaly in which there is a



1851: of note for Foucault (top) and Daguerre.



doubling of the eye lashes in one or both of the eyes.

Those who, on our ever warmer planet, entertain nostalgia for a colder past, may want to notice the future Lord Kelvin's introduction of the absolute temperature scale named after him and the news (for such it was in 1851) that all motion ceases at its zero.

1801 (2.0 ¢)

We face many choices for double-centennial observance, from the heavens (Giuseppe Piazzi's discovery of the first asteroid, Ceres, on 1 January 1801) to the Earth (Franz Carl Achard's invention of the sugar-beet industry). Again among higher things, there was Carl Friedrich Gauss's *Disquisitiones arithmeticae* and the *Histoire céleste française* of that great survivor, Joseph-Jérôme Lalande ("I am an oilskin for insults and a sponge for praise"), which gave the positions of stars in the firmament of French science; and, among lower things, Jean-Baptiste Lamarck's *Système des animaux sans vertèbres* (*System of Invertebrate Animals*), Robert Fulton's submarine, and the discovery of vanadium by Andrés Manuel del Río.

The most important minute in science 200 years ago came in a lecture by Thomas Young, then just appointed professor of natural philosophy at the almost new Royal Institution, on the nature of light: "When two undulations, from different origins, coincide either perfectly or very nearly in direction, their joint effect is a combination of the motions belonging to each." Later Young devised his famous double-slit



Ricci's (shown in the top left inset) map of China.

experiment, which anchored the wave picture of light in his time and the wave-particle picture a hundred years later.

1751 (2.5 ¢)

The capital event in science and technology 250 years ago was the publication of the first volume of Denis Diderot's and Jean le Rond d'Alembert's *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts, et des métiers*, which was to run to 17 volumes of text and 11 of plates. Besides offering the latest in the natural sciences and a guide to trades and manufactures, it presented human history and institutions in the critical liberal style supposed to characterize the scientific mind.

While Diderot and d'Alembert wrote in relative comfort in Paris, their countryman Nicolas Louis de Lacaille was mapping the skies at the Cape of Good Hope under miserable conditions. He measured 10,000 stellar positions, 30 times as many as recorded in the only earlier survey, by Edmond Halley, in 1676–78 (3.25 ¢).

Electricity was in the air and elsewhere in 1751. Michel Adanson studied the jolt of the torpedo fish and — a taste of things to come — Johann Georg Sulzer began reporting his sensations, both pleasant and unpleasant, to the Berlin Academy of Sciences. The disagreeable sensations included the peculiar tang of dissimilar metals, lead and silver, brought together in the mouth. Both the electrical organs of the fish and the odd observations of Sulzer were to help guide Alessandro Volta towards the invention of the electric battery (1800, published in 1801, 2.0 ¢). In the interim, most investigators of electricity followed the ideas that Benjamin Franklin had sketched out in letters published in 1751 as *Experiments and Observations in Electricity, Made at Philadelphia in America*. The booklet introduced the concepts of positive and negative charge, and the procedure for drawing thunderbolts from clouds, which Franklin had been developing for several years at the edge of the civilized world.

1701 (3.0 ¢)

While discoursing of America, we observe that Yale University is celebrating the three hundredth anniversary of its foundation this year.

1651 (3.5 ¢)

William Harvey declared, 350 years ago, in his account of the developing embryo *De generatione animalium*, “*omne vivum ex ovo*” — “everything comes from an egg” — which was scant consolation to a *vivum* hatched in a country with a weak constitution, where, in the coeval words of Thomas Hobbes's *Leviathan*, it would lead a life that was “nasty, brutish, and short”.

Another item that began life in 1651 was Giambattista Riccioli's *Almagestum novum*, which non-mathematicians might think nasty and brutish, but no one thinks short. Riccioli, a Jesuit, was the Catholic Church's point man against Copernicanism. Nonetheless, his *almagest* contains perhaps the best early account of the work of Johannes Kepler and an invaluable inventory of the astronomical knowledge of the day, as well as 49 arguments for and 77 against the motion of the Earth.

1601 (4.0 ¢)

Connoisseurs of Jesuit mathematicians will know that 400 years ago Matteo Ricci, already for some years a missionary in China, moved to the imperial court in Peking. He rendered much of Western exact science into Chinese, notably the first six books of Euclid's *Elements*, and mapped the empire onto a scroll almost two metres tall and over four metres wide.

Pierre de Fermat, over whose last theorem generations of calculators were to crack their heads, drew his first breath in 1601. He was to anticipate analytical geometry and the differential calculus; but lacking the urge to

be an author, he perished without publishing, and left the chore to others.

1551 (4.5 ¢)

Mathematicians could have a perpetual party this year. Georg Joachim Rheticus, who had written the first account of Copernicus's system and arranged for the printing of Copernicus's masterpiece, published his own *Canon doctrinae triangulorum*, the first table to give all six trigonometric functions and a forerunner of his immense and posthumously published *Opus palatinum de triangulis* (*The Palatine Work on Triangles*) which gave their values to ten places at ten-second intervals.

Philip Melanchthon had appointed two professors of mathematics at the University of Wittenberg in 1536: Rheticus for the lower branches and Erasmus Reinhold (born 1511, 4.9 ¢), for the upper. Reinhold undertook what Kepler later called the “huge and disagreeable task” of calculating ephemerides from Copernicus's models of planetary motion, which Reinhold regarded as useful fictions. His *Tabulae prutenicae* (1551) brought the fictions to the working astronomer, with the odd consequence that two Protestant professors working at Luther's university became the agents of the Catholic canon who overturned the world system.

One of Rheticus's fellow students at the University of Zurich and, later, his teacher in medicine, was Konrad von Gesner, who in 1551 published the first of two volumes of his *Opera botanica* (second volume 1571, 4.3 ¢), the first treatise to emphasize floral structure as a key to classification. The two volumes contain almost 1,500 plates, most of which Gesner drew.

1501 (5.0 ¢)

While botanizing, we must mention the birth 500 years ago of Leonhard Fuchs, best



Happy birthday: Yale University celebrates its three-hundredth anniversary this year.

known now as the eponym of the fuchsia. Girolamo Cardano — doctor, mathematician, astrologer, autobiographer and egotist — also qualifies for, and would have appreciated, a half-millennial anniversary.

1451 (5.5 ¢)

Had the world not exhausted itself over the five-hundredth anniversary of the European discovery of America nine years ago, it might have made something this year of the fact that Christopher Columbus was born 550 years ago, as was the man who, by cartographic error, gave his name to Columbus's discovery, Amerigo Vespucci.

1101 (9.0 ¢)

There should be a wingding down at the King's Arm this year. It has been 900 years since King Henry I decreed that the distance between his shoulder and his furthest finger tip would thenceforth be the English yard.

1001 (10.0 ¢)

It deserves repeating that 1,000 years ago, or thereabouts, Leif Erikson reached the coast of Labrador and may have sailed as far south as New York.

901 (11.0 ¢)

Our earliest eligible anniversary is the eleventh centenary of the death of Thabit ibn Qurra, one of the greatest and most prolific of the mathematicians who wrote in Arabic. He perfected his arithmetic as a money changer, gained preferment from the mathematics-minded sons of a bandit, and helped to pave the way to analytic geometry, spherical trigonometry and non-Euclidean geometry. He also wrote on logic, psychology, ethics, Plato's *Republic* and Syriac grammar. Human energy may have declined since the end of the first millennium.

1951 (0.5 ¢); 1951 ± 25 (0.25, 0.75 ¢)

In keeping with the anniversary of the year, we confine our 20th-century recommendations to people who won Nobel prizes at 25-year intervals beginning with the first awards. We have already mentioned the 1901 winners: two Germans and a Dutchman working in Germany. The prizes in 1926 also went to northern Europeans: two Germans (James Franck and Gustav Hertz) received the unattributed physics prize of 1925 (the academy can withhold a prize one year and award it the next); a Frenchman (Jean Perrin), that for 1926; an Austrian working in Germany (Richard Adolf Zsigmondy), the unattributed chemistry prize for 1925; and the Swede Theodor Svedberg, the 1926 prize in chemistry. The Karolinska Institute gave its prize in physiology/medicine for 1926 to a Dane (Johannes Andreas Grib Fibiger). Thus European hegemony.

A century after their alarming showing as inventors at the Crystal Palace in 1851,



Go west: Leif Erikson arrived in what were to become the Americas about 1,000 years ago.

the Americans began to be the favourites of the judges in Stockholm. The prizes of 1951 went to two Americans (Edwin Mattison McMillan (died 1991, 0.1 ¢) and Glenn Theodore Seaborg), a South African working in the United States (Max Theiler), an Englishman (John Cockcroft), and an Irishman (Ernest Thomas Sinton Walton). In 1976, the Yankees made a clean sweep: Burton Richter and Samuel Ting in physics, William Lipscomb in chemistry, Baruch Blumberg and D. Carleton Gajdusek in physiology/medicine.

Except for the award to Jean Perrin for his demonstration of the atomic nature of matter, all these physics prizes were awarded for work with particle accelerators. Röntgen made X-rays by slamming electrons into the glass wall of the cathode-ray tube. Franck and Hertz found what others recognized to be quantized energy exchanges by bombarding gas molecules with electrons. Cockcroft and Walton were the first to disintegrate atomic nuclei with a particle beam created by a machine. Richter used a storage ring at the Stanford Linear Accelerator Center to bring electrons and positrons into violent collision whereas Ting used the synchrotron at Brookhaven to crash protons into beryllium, thus creating two new particles recognized as one (the J/ψ) and a case for the existence of quarks with charm. It is a violent and peculiar world.

The chemistry prizes show the greater variety of a science without a dominating sector. Van't Hoff won for discovering laws of chemical dynamics; Zsigmondy, for work on colloidal suspensions; Svedberg, for study of sedimentation; McMillan and Seaborg, for detection of transuranic elements, including the notorious plutonium; and Lipscomb, for the structure of boranes and associated problems of chemical bonding. The anniversary, intent as always to find or feign

connections, cannot help but think that the Nobel committees on chemistry and physics sometimes followed one another's lead. Perrin's prizewinning work, like Zsigmondy's and Svedberg's, centred on colloidal suspensions. McMillan and Seaborg, and Cockcroft and Walton, both used accelerators to transform elements: the Berkeley 60-inch cyclotron in the first case, a purpose-built proton accelerator in the second. But whereas Cockcroft and Walton did physics at the bottom end of the periodic table by reducing lithium to alpha particles, McMillan and Seaborg did chemistry at the top by synthesizing plutonium from uranium.

The medicine/physiology awards show a preoccupation with ever-decreasing units of disease causation, from Fibiger's triumphant demonstration (now discarded knowledge) that nematodes can cause carcinomas in rodents, to the viruses of yellow fever (Theiler) and hepatitis (Blumberg) and the first inkling that even proteins can be dangerous, as Gajdusek found with the brain disease kuru.

Had the Karolinska committee followed Nobel's wishes, and instead selected only work done in the year preceding the award, it might have noticed Karl Kundratitz's demonstration of the infectivity of herpes zoster (1925), the establishment of the association between smoking and lung cancer by Austin Bradford Hill and Richard Doll in Britain, and Ernst Wynder and Evarts Graham in the United States (1950), or the production of monoclonal antibodies by César Milstein and Georges Köhler (1975) (\$1984). Maybe Nobel knew what he was doing. ■

J. L. Heilbron is at Worcester College, Oxford OX1 2HB, UK. W. F. Bynum is at the Wellcome Trust Centre for the History of Medicine at University College London, 24 Eversholt Street, London NW1 1AD, UK.



After Milstein: the Nobel for monoclonal antibodies was finally awarded in 1984.

The melting-pot that is ALife

An initiation into the world of artificial life by one of its architects.

Creation: Life and How to Make It
by Steve Grand
Weidenfeld & Nicolson: 2000. 230 pp. £18.99
John L. Casti

The credo of 'artificial life' (ALife) aficionados is that life is about function, not form. Or, put another way, what separates the living from the dead is not a matter of matter but resides in patterns of information. Here on Earth, these patterns of information happen to manifest themselves in various arrangements of chemical elements such as carbon, oxygen, phosphorus and nitrogen. But it could have been otherwise, say the ALifers, and life can just as easily be created in bits and bytes inside structures of silicon as it was created in the chemical compounds of the primeval seas. Steve Grand's book is a gentle — but thorough — argument of this case.

Grand certainly possesses the credentials for the job, having been the architect and chief programmer of *Creatures*, the first computer game to use real artificial life, which has nearly a million users worldwide. In addition, Grand was nominated by *The Sunday Times* as one of the "Brains Behind the 21st Century" and was awarded an OBE in the Millennium Honours list. So when someone like Grand talks, people listen.

Grand starts his lessons in life by stating "the most important law of nature": things that persist, persist; things that don't, don't. This obvious tautology leads to the rather profound — but equally obvious — fact that the Universe looks the way it does and not some other way because some things persist and others do not — including life. In this framework, matter becomes a disturbance in space, not a superimposition of something on it, and thus both matter and energy as well as things such as genes, minds and ideas can be thought of as being made out of the same sort of "non-stuff". Grand rightly says that this point of view is a much-needed antidote to post-Newtonian physical science, which tries to explain the Universe purely in material terms.

Using consciousness to illustrate the point, Grand argues that this cannot be a property of matter, but only a property of certain configurations of matter. The objective of the book is then to consider how we might configure large numbers of components to create life — and maybe even consciousness. Of course, we must be very explicit about what type of components we need, how they should be configured and how many of them are required for the job. In essence, the answers to these questions constitute the balance of the book.



Grand design: Steve Grand with artificial chimpanzee, Lucy, his first robot life-form.

Following these preliminaries, Grand focuses on several of the staples in the system theorist's tool-chest: hierarchies, stability, self-regulation, feedback loops, and what is perhaps the single most important system-theoretic concept of them all, emergence. More than any other single notion, it is emergence that accounts for how a collection of primitive individual objects such as a neuron or a cell can form a population possessing properties that no single object in the group displays. A graphic example of this is the explosive, nitroglycerin, formed by mixing nitric acid and glycerol. You could put either of these liquids in a bottle and shake it until the Sun dies and nothing very exciting would happen. But if the same experiment is done with a mixture of the two, there are some explosive — and immediate — consequences. So the explosiveness of the mixture resides in the interaction of the components, not in the components taken in isolation. This phenomenon pervades complex systems of every type.

Midway through the book Grand shifts gear and begins a lengthy and illuminating account of the creation of software objects that he claims can be fairly described as 'alive'. The basis for this rather grand claim (no pun intended) is that we can create simulated objects that embody the cybernetic properties noted above from which life emerged in the natural world. But, as Grand properly notes, it's necessary to ensure that

this emergent phenomenon is not hard-wired into the code but can be passed on from generation to generation and can modify itself in order to persist on longer timescales, as the environment changes.

So the creatures brought into existence by this procedure will be fully alive and intelligent only if their future rests in their own hands. And to give them this kind of autonomy we must give up direct control of their design. Grand spends the next 50 pages or so (one-fifth of the entire book) outlining how this can be done.

Having very cogently and entertainingly presented his case, Grand concludes with a few ruminations and reflections on popular myths surrounding artificial intelligence and ALife. These include the feelings often expressed by laypeople and those of a humanistic bent that such artificial intelligences are something to be angry about or feared, or even that they cannot possibly exist. As Grand states it, such folks seem to lump the idea of intelligent living machines alongside such things as nuclear power and genetically modified food. I think the best part of the book is Grand's refutation of these irrational fears.

On balance, *Creation* is an honourable and, for the most part, successful attempt to convey the flavour of the ALife movement to the uninitiated. Given the book's target audience, it's probably inevitable that most of it centres on topics and issues familiar to people working in the systems area. Personally, I

book reviews

would have been happier if the publisher and/or author had added an extra 50 pages or so to what is a rather short book and included a few more details of exactly how one creates a game such as *Creatures*, with explicit indications of where the various notions such as self-organization and adaptation come into play in the *Creatures* world. But this is a minor quibble.

If you've heard about ALife but aren't quite sure what it is or where it's going, Grand's book is an excellent place to enter one of the more exciting areas of twenty-first-century science. ■

John L. Casti is at the Santa Fe Institute,
1399 Hyde Park Road, Santa Fe, New Mexico
87501, USA.

The man with faith in the unseeable

Boltzmann's Atom: The Great Debate That Launched a Revolution in Physics
by David Lindley

Free Press: 2000. 256 pp. \$24

Jacqueline Reynolds and
Charles Tanford

The Zentralfriedhof on the edge of Vienna is a place of pilgrimage for many tourists. The city's largest burial ground, it has a special section, the Ehrengrüber, where Beethoven, Schubert, Brahms, Strauss and various Viennese *Bürgermeister* are buried. Among these luminaries lies one of the great heroes of modern-day physics, Ludwig Boltzmann. His tombstone is inscribed with the simple equation, $S = k \log W$, which may bemuse visiting music lovers but is a fitting tribute to the man who linked the phenomenological science of thermodynamics to the concept of atoms and molecules.

The true Boltzmann enthusiast (which it must be admitted includes the reviewers) would probably also visit the city of Graz, where Boltzmann spent his most productive years. There they could view the magnificent display of physical instruments he and his colleagues built for instructional purposes. Few, if any, university lecturers today would put such enormous effort into teaching — after all, it is research that brings in the money.

Unfortunately, there are no definitive biographies of Boltzmann because of the scarcity of personal records about his life outside science. There is, however, a wealth of information about his scientific activities and his interactions with the worldwide scientific community. David Lindley has drawn on these latter sources to produce, not a biography of the man, but rather a biography of an idea — the debate about the



Atomic force: Boltzmann linked thermodynamics to the concept of atoms and molecules.

existence of atoms that raged in Europe in the latter years of the nineteenth century. All the familiar characters — Mach, Ostwald, Planck, Stefan, Maxwell, Einstein, Gibbs to name only a few — come alive for the reader through entertaining explanations of their work and ideas coupled to insightful glimpses of their personalities.

But the real core of this book is atomic theory as viewed by the great and the good from about 1865 to 1906, when the unfortunate (and certainly unstable) Boltzmann committed suicide. By this time, all but the most diehard of the “energeticists” had accepted the reality of atoms. The philosophical differences between those who believed in the unseen atoms and those who didn't are an important part of the story, as are the cultural differences among the adversaries. In England there was little or no resistance to atomic theory other than from the then aged and crotchety Lord Kelvin, who in 1900 is described in a letter from Josef Nabl to Stefan Meyer in Cambridge as “a silly old idiot”. But in the rest of Europe, many scientists were still in thrall to a Germanic philosophy opposed to theorizing about the ‘unseeable’. They believed that all physics should deal only in the realm of the observable, with ultimate truth lying in empirical laws that described such phenomena. Had Boltzmann lived in England rather than Austria, he would certainly have had an easier time.

Lindley gives a superb account of the protracted attempt to explain the second law of thermodynamics in statistical terms, first by James Maxwell and then by Boltzmann. Both men envisaged matter as consisting of atoms obeying classical Newtonian laws. The confusion engendered by Boltzmann's H-theorem, and his attempts to explain away the anomalies in its implications that others discovered, is one of the best exposés of the sub-

ject we have read (albeit somewhat repetitious at times). The only slight flaw in Lindley's rendering of the debate is his failure to point out that those who opposed the statistical approach did have one reasonable objection — that classical statistical mechanics (as it came to be called by J. Willard Gibbs in 1901) could not explain the specific heats of gases. This problem had to await the advent of quantum theory in the early 1900s.

Another strange omission is from the early part of the book, in which the history of the idea of atoms is described in great detail. How can one go from the early Greeks to Maxwell's kinetic theory of gases and ignore the “father of the atom”, John Dalton? Perhaps the explanation lies in the age-old antagonism between physicists and chemists: Lindley was trained in the former discipline whereas Dalton, of course, is claimed by the chemists as one of their own.

It is a great temptation to shower superlatives on a delightful account of one's favourite topic, and here we plead some personal bias. But, with that caveat, we enthusiastically recommend this book, not only to chemists and physicists, but to any scientist with an interest in the history of the subject. Perhaps that category should include us all, since we have grown up in a world firmly founded on the certainty of the chemical/physical atom. ■

Jacqueline Reynolds and Charles Tanford are at
Tarlswood, Back Lane, Easingwold, North
Yorkshire YO6 3BG, UK.

For the birds

Bird Nests and Construction Behaviour

by Michael Hansell

Cambridge University Press: 2000. 280 pp.
\$80, £50

Tore Slagsvold

For many of us, the discovery of a bird's nest when we were young created a life-long interest in birds and nature. The nest was a revelation, its rounded shape and feather lining making it a perfect bed for the eggs and a perfect cradle for the small chicks, rocking them to sleep in the soft breeze. Later, we learnt that nests were subject to natural selection, their final solution being a compromise between various demands. For example, it must be small enough to avoid detection by a predator and to fit tightly around the incubating parent, but large enough to provide space for a big brood of chicks scrambling for food. And it must be built with materials possessing good insulation properties, able to keep out moisture and to dry quickly after rain, and with the strength to stay put in the wind.



Nesting niceties (clockwise from top right): house martin, hammerkop, song thrush and village weaver.

All these aspects are included in Mike Hansell's well-written book, which provides an excellent overview. Also included are many references covering the important discoveries made since the classical books, on animal constructions, *Animal Architecture and Building Behaviour* (Longman) by Hansell, and on bird nests, *Nest Building and Bird Behavior* (Princeton University Press) by N. E. Collias and E. C. Collias, were published in 1984.

Of interest to anyone with a passion for evolutionary biology, the book describes fascinating works of engineering and architecture, ranging from how the birds' ancestors — dinosaurs — made their nests, to how a male black wheatear carries more than two kilograms of stones to impress a female.

One way to reduce the cost of nest building is to steal materials from a neighbour's nest, as many colonial birds do. Another solution is to reuse a vacated nest, although this has the drawback of a possible legacy of blood-sucking bugs. The latter problem can be alleviated, however, by the addition of green leaves possessing active agents against mites and by importing spiders to kill flies.

Hansell places all this knowledge into a framework of modern behavioural ecology, raising questions such as whether nest size is a constraint on optimal brood size, and whether selection of nest sites and nesting

materials is involved in speciation. He also analyses the nests of more than 500 specimens from museum collections. His detailed categorization will probably serve as a base reference for years to come.

Is it necessary to have a big brain to build a nest? No, says Hansell. Some learning may be involved, as is seen in weavers, where experienced birds perform better than younger ones. But the great variety of sophistication of bird-nest construction is achieved with a surprisingly small repertoire of rather stereotyped behaviours and by using particular nesting materials, such as silk from insects and spiders, and vine tendrils or spiny lichens that act as 'Velcro', locking together easily.

Nest building can be likened to tool use by birds, says Hansell. Crows, which have relatively large brains, may even use narrow sticks to pick out insects during foraging. But birds are also social animals, and perhaps, as with humans, the evolution of increased brain size has been driven by complex social demands rather than by the use of tools or the construction of nests.

In any case, bird nests are beautiful and serve their purpose well. The same can also be said of this book. ■

Tore Slagsvold is in the Department of Biology, University of Oslo, PO Box 1050, Blindern, N-0316 Oslo, Norway.

book reviews

Virology in all its guises

The Invisible Enemy: A Natural History of Viruses

by Dorothy H. Crawford
Oxford University Press: 2000. 288 pp.
£14.99, \$25

Albert D. M. E. Osterhaus

Dorothy Crawford's book portrays viruses as unseen enemies, with many tricks up their sleeve, and this approach leads to a fascinating story of their natural history. For the educated layperson, the book reads like a novel, despite starting with a textbook approach, explaining what a virus is and how it takes over the normal metabolism of living cells. Once the author has completed these basics, though, the book builds to a climax, describing as it develops the plethora of mechanisms by which viruses may be involved in the development of cancer.

Most of the topics dealt with, including stories of 'everyday' virus infections such as the common cold, influenza, varicella and shingles, are illustrated with real examples from medical rather than veterinary practice, a bias that reflects the author's own research interests. Of the examples, most come from Britain, showcasing the role of British pioneers in virology.

The account starts with the important contributions of Edward Jenner, who introduced the principle of vaccination. It continues with the story of Christopher Andrewes' and David Tyrrell's invaluable work at the Common Cold Unit in Salisbury and ends with Anthony Epstein, who identified the role of Epstein-Barr virus in Burkitt's lymphoma.

The darker or more controversial sides of the history of virological research are also dealt with. For example, Crawford explores the controversy between Luc Montagnier and Robert Gallo over their roles in the discovery of HIV-1 as the cause of AIDS, and she highlights Peter Duesberg's currently untenable theory denying that HIV-1 is the sole cause of the disease. Furthermore, the so-far unsuccessful attempts to identify viral causes for other unresolved human diseases or syndromes, such as chronic fatigue syndrome, Gulf War syndrome and multiple sclerosis, present the reader with a glimpse of the challenges of modern virology.

From a historical perspective, the author explains the important role of vaccines in eliminating or eradicating viral diseases such as smallpox, polio and measles. Then, bringing the story right up to date, she discusses the increasing importance of antiviral drugs, using AIDS to illustrate the problems associated with their development and use.

The book ends by turning the tables and speculating on ways in which the invisible

book reviews

enemy may have a future use as a carrier of therapeutic genes for combating cancer and other scourges of mankind. Altogether, this fascinating book provides a rapid and accessible introduction to modern virology. ■

Albert D. M. E. Osterhaus is in the Department of Virology, Erasmus University, Room EE17-26, PO Box 1738, 3000 DR Rotterdam, The Netherlands.

The rocky road to dating the Earth

The Dating Game: One Man's Search for the Age of the Earth

by Cherry Lewis

Cambridge University Press: 2000. 253 pp. £17.95, \$24.95

Douglas Palmer

I doubt that many Regius professors and Fellows of the Royal Society today have had to interrupt their academic career to earn their living as a shopkeeper. This necessity fell to Arthur Holmes, one of twentieth-century Britain's foremost geologists. And although such mundane commercial necessity certainly held back his research into the radiometric dating of rocks, it does not seem to have embittered him in any way. In this book, Holmes emerges as a much-admired man, a famously good advocate for geology in Britain and a remarkably persistent researcher.

Much of the book is a biography detailing Holmes' experiences during the first decade of the twentieth century, as a schoolboy and a student of physics at the Royal College of Science in London. This period saw the public debate that led to the final escape from the straitjacket of Lord Kelvin's calculation that the world was only some 20 million years old. Cherry Lewis broadens her scope to explain the background to the debate and how Kelvin was working on a model of the Earth's evolution in which thermal diffusion acted to create a progressively thickening crust on an initially molten Earth.

Kelvin "argued ... that temperatures within the Earth increased with depth. From this he deduced that the Earth was cooling ... from a molten globe." Knowing the temperature at which rocks become molten and the rate at which they cool, he calculated the time taken for the Earth to cool and therefore its age. But, as we now know, this was a serious underestimate.

One of the early workers in the post-Kelvin era was Ernest Rutherford. In 1902, using the decay of radioactive elements, he measured the age of formation of a piece of uranium ore at 700 million years. But his method was flawed because it relied on measuring the gas helium, which can easily escape from rock. This problem with Rutherford's calculations was first recog-

nized over the next few years by Robert Strutt, professor of physics at the Royal College. And in 1907, the American chemist Bertram Boltwood showed that uranium-rich rocks contained large amounts of lead along with helium. Boltwood postulated that the lead was the stable end-product of the decay chain from uranium.

Under the inspired tutelage of Strutt and William Watts at the Royal College, Holmes combined his studies of physics with that of geology. Final-year undergraduates at the Royal College were encouraged to conduct an original piece of research. And in 1910, according to Lewis, "Holmes realised ... it should be possible to obtain an age by measuring the amount of lead present in the mineral, rather than the amount of helium". Holmes eventually calculated the age of a Norwegian rock as 370 million years. As the rock was known to have originated in the Devonian geological system, he thus provided the first date for that system. He also recalculated some of Boltzmann's published data and arranged them to produce a geological timescale. He was to improve on this continuously for the rest of his professional life.

Life for a student in London was as difficult financially then as it is now, and Holmes ran into debt and could not even afford the £5 (US\$7) membership for the Geological Society. For the first, but not the last, time he had to get a job to pay his way. He spent a year prospecting for minerals in Mozambique, and returned — having survived blackwater fever and malaria — with a desire to develop a geological timescale based on radiometric dating and £89 7s 3d profit for himself, but not much for the mining company.

In 1920, by then married with a child, Holmes was trying to survive on £150 a year as a demonstrator at the Royal College, which by then had been transformed into Imperial College. Another foreign adventure was necessary, this time looking for oil in Burma. As unsuccessful as the last one, it was also personally disastrous, as his son contracted dysentery and died there. Returning to Gateshead in 1922, without a job or prospects, Holmes joined forces with his wife's cousin and opened a shop in Newcastle, trading in "oriental crafts".

Holmes' luck changed in 1924, when he got a job as Reader in the newly founded department of geology at Durham University. He stayed there until 1943, spending most of his time building up the department and providing service teaching for the other sciences. But he also managed to continue his work of radiometric dating and extended Alfred Wegener's work on continental drift into a new hypothesis based on sub-crustal



Inspiration: Holmes' pioneering work stimulated generations of geologists.

convection. The role of mantle convection as a mechanism for plate tectonics driven by the circulation of heat from the Earth's interior is now virtually taken for granted. Holmes' speculation was published in 1931, the same year in which he met a London-based academic geologist Doris Reynolds.

But Holmes' personal life led to problems in the ecclesiastically dominated Durham. According to Lewis, Holmes' wife Maggie "did not enjoy university life", and although another child had been born, the marriage did not fulfil either partner. In 1933, Doris Reynolds took up a lectureship in Durham, and rumours spread about her relationship with Holmes. Although Maggie died in 1938 and Holmes married Doris, the university authorities showed their displeasure by renewing Doris's contract for only one year. But Holmes' appointment in 1943 to the prestigious Regius professorship of geology in Edinburgh brought security and freedom.

The publication of the first edition of *Principles of Physical Geology* in 1944 secured Holmes' reputation as the pre-eminent proselytizer of the subject in the British Isles. The book enthused several generations of geologists who built on the work of Holmes and other stalwarts of his generation. Much of the lost reputation for geological research in Britain was thereby regained. In 1953, Claire Patterson was the first to finally date the age of the Earth at about 4,500 million years old. She wrote to Holmes: "I wish to reiterate my personal indebtedness to your pioneering work in this field. It was outstandingly inspiring and ingenious."

Cherry Lewis, a mature student of geology who subsequently worked in research and in the oil industry, is no historian of science. But she has a good understanding of the technical details of the subject and gives a sympathetic portrayal of Holmes' struggle to match economic necessity with research ambition.

One of the book's pervading themes is the pathetic state of British universities, particularly their science departments, during the first half of the twentieth century. Even as Regius professor, Holmes had to make a special application to the university for £74 8s for the purchase of an electronic Marchant calculating machine.

To conclude, *The Dating Game* is a welcome portrait of a gifted British scientist whose abilities were often stymied by lack of funds and resources. ■

Douglas Palmer is at 31 Mawson Road, Cambridge CB1 2DZ, UK. His most recent book, *Neanderthal*, has just been published by Channel 4 Books.

Genius loci

The twentieth century was made in Budapest.

Vaclav Smil

The ancient Romans had a term for it — *genius loci* — and history is not short of astounding, seemingly inexplicable concatenations of creative talent. Florence in the first decade of the sixteenth century is perhaps the unmatched example: anyone idling on the Piazza della Signoria for a few days could have bumped into Leonardo da Vinci, Raphael, Michelangelo and Botticelli. Other well-known efflorescences of artistic creativity include Joseph II's Vienna in the 1780s, where one could have met C. W. Gluck, Haydn and Mozart in the same room. Or, eleven decades later, in *fin de siècle* Paris one could read the most recent instalment of Émile Zola's Rougon-Macquart cycle, before seeing Claude Monet's latest canvases from Giverny, and then strolling along to a performance of Claude Debussy's *Prélude à l'après-midi d'un faune* in the evening.

But it is not just today's young adults — who probably view Silicon Valley as the centre of the creative world — who would be unaware that an improbable number of scientific greats were born in Budapest in the decade between 1898 and 1908. Between them, this group were responsible for some of the twentieth century's most decisive scientific advances and, consequently, some of its fundamental strategic and political transformations.

Leo Szilard, a physicist who both studied and worked with Einstein and who, together with Enrico Fermi, patented the first nuclear fission reactor, was born there in 1898. In the summer of 1939, Szilard and Eugene Wigner, born in the city in 1902, persuaded Einstein to sign the famous letter to President Franklin Roosevelt that led to the Manhattan Project. Dennis Gabor, whose research ranged from pioneering work in holography to nuclear fusion, was born in 1900, and John von Neumann three years later.

Von Neumann's prodigious feats of problem-solving during the Second World War — prefigured by his ability to divide eight-digit numbers in his head at the age of six — have been overshadowed by his postwar conception of the stored computer program, the prototypical architecture of modern computers (although when told in 1954 of the idea for FORTRAN, he asked: "Why would you want more than machine language?").

Edward Teller, born in 1908, is the only living member of this group. His fame will always rest on his contribution to the design of America's first thermonuclear weapon,



Blue streak: Turn-of-the-century Budapest produced a plethora of great minds, especially in physics.

and on his later advocacy of antiballistic missile defences.

By pushing the time frame back a bit, and by admitting bright intellects from beyond physics, the Budapest circle must be enlarged — to mention just its most prominent over-achievers — by Theodore von Kármán (1881–1963), a pioneer in aerodynamics and aeronautics whose studies of fluid flows helped to open the era of fast subsonic and supersonic flight; by Albert Szent-Györgi (1893–1986), who, after isolating ascorbic acid (for which he won the Nobel Prize in Physiology for 1937), went on to identify actin and myosin, the proteins responsible for muscle contraction; by Michael Polanyi (1891–1976), who was not just an outstanding physical chemist but also an accomplished economist and philosopher; and by Arthur Koestler (1905–83), a brilliant writer and one of the most incisive chroniclers of the great political and scientific upheavals of the twentieth century.

Besides their birthplace, these men had a number of other things in common. Most of them came from the city's German-speaking Jewish families, but Szent-Györgi was born to a rich land-owning family and Gabor's

father was the director of a mining company. All of them left their birthplace to attend university either in Germany (mostly Berlin and Karlsruhe) or at Zurich's ETH. And all of them ended up either in the United States or the United Kingdom.

But the differences among them are no less remarkable. Three of the group — Szent-Györgi in 1937, Wigner in 1963 and Gabor in 1971 — got Nobel prizes. Szilard, with his myriad of interests, never settled in one place, and his fundamental contributions to modern science are not generally appreciated. Von Kármán, von Neumann and Teller contributed much to the United States' rise to postwar strategic dominance.

No single fact can explain this phenomenon. Budapest was not the only city in the Austro-Hungarian empire brimming with creativity at this time. In the decade before the First World War, intellectuals such as Sigmund Freud, Gustav Mahler and the physicist Ernst Mach worked in Vienna. Meanwhile, Franz Kafka, the painter Alfons Mucha and the poet Rainer Maria Rilke were in Prague, where, in 1911–12, Einstein was developing his general theory of relativity. A number of factors that von Neumann identified as being behind the Budapest phenomenon were present in the other two cities: a multicultural environment, external pressure to succeed, "a feeling of extreme insecurity in the individuals, and the necessity to produce the unusual or else face extinction". But, in the end, only the Budapest group made such an improbable — and incomparable — mark on history. ■

Vaclav Smil is at the University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada.

Von Neumann identified "the necessity to produce the unusual or else face extinction".

Time for gas planets to grow

Jack J. Lissauer

Giant planets like Jupiter need a large reservoir of gas to grow to full size. New observations indicate that such planetary nurseries last twice as long as previously thought.

Until now, astronomers have believed that giant gas planets, such as Jupiter and Saturn, must form in a few million years or less — quite fast in astronomical terms. On page 60 of this issue, however, Thi *et al.*¹ report evidence to the contrary. They have detected substantial amounts of molecular hydrogen (H_2) gas around three youngish nearby stars in spectroscopic measurements taken by the Infrared Space Observatory spacecraft.

These stars were already known to possess dust disks, so why is it surprising that they are also surrounded by molecular hydrogen, the most common of all the elements, and why should anyone not specializing in stellar environments care? We care because hydrogen is the main ingredient needed to make the giant planets in our own Solar System, and its discovery around other stars means that giant-planet formation may be more common than previously thought. The result is surprising to astronomers because very little carbon monoxide gas orbits these stars. Although it is far less abundant, carbon monoxide is much easier to observe than hydrogen, and has even been detected around a very distant quasar, as described by Papadopoulos *et al.*² on page 58 of this issue.

Models of planet formation have been developed primarily to explain the existence of planets and smaller bodies within our Solar System³. The major planets have almost circular orbits in roughly the same plane, suggesting that they formed from a disk of material orbiting the Sun. The inner planets, moons and asteroids have a rocky composition, whereas most moons and small bodies beyond the asteroid belt are rich in ice. All of these bodies grew by condensing the material around them. The variation in composition implies that the temperature of the disk material decreased away from the centre of the Solar System.

The four jovian planets (Jupiter, Saturn, Uranus and Neptune) have large masses but low densities, so they must mainly be composed of light materials, such as hydrogen and helium. The most popular models for the formation of giant planets begin with the accumulation of a core of rock and ice roughly ten times the mass of Earth. The core grows into a giant planet by gravitationally accumulating hydrogen and helium gas from the surrounding protoplanetary disk. These models gener-

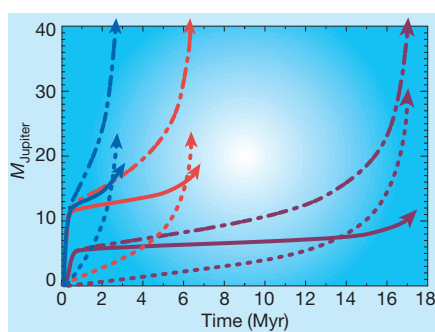


Figure 1 Three models of the growth of Jupiter, calculated as in ref. 10. The vertical scale gives the mass of a growing Jupiter, in units of Earth's present mass. The curves show the gaseous (dotted) and solid (solid) components of Jupiter, as well as the mass of the entire planet (dot-dashed). Initially, the planet accumulates nearly all of the small planetary masses within its gravitational reach. When the planet is massive enough, it contracts and rapidly accretes all nearby gas. The planet's rate of gas accumulation is governed by its ability to radiate away energy that is released by gravitational contraction. The red curves assume that the dust-to-gas ratio within the planet's atmosphere is the same as the interstellar ratio; the blue curve assumes a clearer atmosphere with 2% of the interstellar value, so the planet can radiate energy and accumulate gas faster. The purple curves also assume a low dust atmosphere, but less rock and ice available for the core, so the planet's gravity is weaker. The new observations of Thi *et al.*¹ can accommodate models with a broader range of atmospheric dust abundance and core sizes, some of which predict long formation times for the giant planets (red and purple curves).

ally require a few million years to form a Jupiter-like planet (Fig. 1; blue curve). Although this is rapid compared to the 20–100 million years believed necessary to form Earth-like planets^{3,4}, giant planets require a substantial gas reservoir to complete their growth, whereas rocky planets can keep growing in a gas-free environment like that of our Solar System today. So according to current models of planet formation, these giant planets must accumulate gas fairly rapidly.

Do giant planets form only in unusually long-lived protoplanetary disks, or can they accumulate around most stars? The measurements of Thi *et al.*¹ suggest that gas remains in many protoplanetary disks long

enough to form Jupiter-like planets. Massive dust disks are known to occur around most stars less than 1 million years old, but few stars older than 5 million years possess such disks. Thi *et al.* have now measured the amount of H_2 gas in the smallish dust disks found around three stars between 8 million and 30 million years old. Previous observations of these stars, based on measurements of carbon monoxide gas, suggested that there is surprisingly little gas left around the star, giving a much lower gas-to-dust ratio than found in the interstellar medium (which is mostly hydrogen). But Thi *et al.*'s hydrogen measurements show that these stars have around one Jupiter mass of H_2 in their dust disks, with gas-to-dust ratios similar to that in the interstellar medium. One of the three disks may still have enough gas for the growth of Jupiter-mass planets. The presence of H_2 but not carbon monoxide may be explained by carbon monoxide condensing onto dust grains in icy conditions, or being destroyed by ultraviolet starlight.

More generally, Thi *et al.*'s results suggest that dust is a good indicator of hydrogen in disks around young stars. The discovery of larger amounts of gas in the disks of older systems suggests that jovian planets can form on timescales of up to 10–20 million years within some disks (Fig. 1; purple curve). This means that the formation of giant planets is likely to be fairly common, at least around isolated Sun-like stars. The situation will be different in a star-forming cloud that is producing luminous massive stars, whose bright ultraviolet radiation could destroy nearby protoplanetary disks⁵.

What else can be learned from observations of this kind? Since 1995, giant planets have been discovered around 50 other stars. Although the masses of these planets are all Jupiter-like, their orbits are not. Jupiter is more than five times the Earth–Sun distance from our star, whereas most known extrasolar planets are less than two Earth–Sun distances from theirs. Extrasolar planets at Jupiter's distance are difficult to detect, so they may be quite abundant. Nonetheless, the observed orbits are hard to explain because it is unlikely that such large planets formed so close to their star.

Further measurements of hydrogen on dust disks around a wide range of stars (both young and old) may be the key to new insights in this area. For example, the man-

O. HUBICKY



100 YEARS AGO

Science is cosmopolitan. Electricity abolishes time and envelops both hemispheres with a new idea as soon as it has emerged from the brain of the Thinker. Mechanics, by its space-annihilating power, has reduced the surface of the planet to such an extent that the human race now possesses the advantage of dwelling, as it were, on a tiny satellite. Both these agencies, then, combine to facilitate a rapid exchange of new ideas and commodities, as well as of those who are interested in them in whatever capacity. These considerations indicate some of the most momentous changes which have occurred in the world's history since the last century dawned ... The enormous and unprecedented progress in science during the last century has brought about a perfectly new state of things, in which the "struggle for existence" which Darwin studied in relation to organic forms is now seen, for the first time, to apply to organised communities, not when at war with each other, but when engaged in peaceful commercial strife. It is a struggle in which the fittest to survive is no longer indicated by his valour and muscle and powers of endurance, but by those qualities in which the most successful differs most from the rest.

From *Nature* 3 January 1901.

50 YEARS AGO

During the past hundred years or so a variety of techniques has been devised for transmitting messages electrically from one point to another. It is only of recent years, however ... that means have been provided for assessing quantitatively the commodity which is transmitted, namely, the 'information' content of messages, and of determining the extent to which existing techniques fall short of what may be attainable. This recent work proves to have a significance beyond the sphere of electrical communications. A new branch of science is emerging which reveals and clarifies connexions between previously largely unrelated fields of research concerned with different aspects of the processes by which living organisms — in particular man — collect, classify, convert and transmit information. A confluence of different fields of investigation is, of course, no new phenomenon in the history of science, but the wide recognition of a new connecting link can seldom have been so rapid as in the present case.

From *Nature* 6 January 1951.

ner in which gas is removed from a proto-planetary disk could have as much influence on the ultimate configuration of the planetary system as does the lifetime of the disk. A planet gravitationally tugs surrounding disk material, and this interaction can alter planetary orbits substantially. Although the possibility of significant planetary migration was predicted more than two decades ago⁶, this type of interaction was largely ignored, because theory suggested that a planet would move faster as it approached a star. Planets capable of migrating a significant distance were therefore expected to spiral inwards to a hot death.

The discovery of the first Jupiter-mass planet orbiting at only one-twentieth of the Earth–Sun distance from its star⁷ (with an orbital period of 4.2 days) led to the suggestion that planetary migration could be stopped very close to the star. This could happen either because the planet enters a gas-free orbit, cleared by magnetic processes close to the star, or because the planet experiences counterbalancing forces resulting from tidal motion on the star that is induced by the planet⁸. But these models do not account for the many giant planets subsequently discovered with intermediate orbital periods ranging from a few weeks to a few

years⁹. Such planets are still too close to their star to have grown *in situ*¹⁰, yet too far away for the proposed stopping mechanisms to operate. Might these giant planets have been migrating inwards only to be stranded as the star cleared away disk material from the inside outwards? The Space Infrared Telescope Facility¹¹, to be launched by NASA in 2002, will be able to observe disks containing hydrogen at much higher resolution, so hundreds of nurseries will soon be available to test this and other ideas of giant-planet formation.

Jack J. Lissauer is at the NASA Ames Research Center, Moffett Field, California 94035, USA.

e-mail: lissauer@ringside.arc.nasa.gov

1. Thi, W. F. *et al. Nature* **409**, 60–63 (2001).
2. Papadopoulos, P., Ivison, R., Carilli, C. & Lewis, G. *Nature* **409**, 58–60 (2001).
3. Safronov, V. S. (Nauka, Moscow, 1969) [in Russian]. Transl.: *Evolution of the Protoplanetary Cloud and Formation of the Earth and Planets* (NASA TTF-677, 1972).
4. Wetherill, G. W. *Annu. Rev. Earth Planet Sci.* **18**, 205–256 (1990).
5. Johnstone, D., Hollenbach, D. & Bally, J. *Astrophys. J.* **499**, 758–776 (1998).
6. Goldreich, P. & Tremaine, S. *Astrophys. J.* **241**, 425–441 (1980).
7. Mayor, M. & Queloz, D. *Nature* **378**, 355–359 (1995).
8. Lin, D. N. C., Bodenheimer, P. & Richardson, D. C. *Nature* **380**, 606–607 (1996).
9. <http://exoplanets.org/>
10. Bodenheimer, P., Hubickyj, O. & Lissauer, J. *J. Icarus* **143**, 2–14 (2000).
11. <http://sirtf.jpl.nasa.gov/>

Neurobiology

Background inhibition to the fore

Ivan Soltesz and Zoltan Nusser

Investigations of a neurotransmitter receptor required for 'background' neuronal inhibition in mice show the importance of such inhibition in keeping neuronal excitability under control.

Why would you put microphones outside a concert hall to record the music inside? All they would pick up is muffled noise, although the sound would vary according to the music's highs and lows. On the other hand, with sufficiently sensitive microphones you could monitor several concert halls simultaneously, and thus keep an ear on the musical life of the whole town. The festival halls of the nervous system are the tiny junctions, called synapses, between nerve cells. And the molecular microphones — neurotransmitter receptors — are indeed found outside synapses, as well as inside them, on the surface of nerve cells^{1,2} (Fig. 1). In fact, there might be more receptors outside synapses than inside³. But despite their abundance, the function of these 'extrasynaptic' molecular microphones has been elusive. Writing on page 89 of this issue, Brickley and colleagues⁴ make a significant contribution to deciphering their importance.

Brickley *et al.* have studied inhibitory neuronal signalling mediated by the neuro-

transmitter γ -amino-butyric acid (GABA). Inhibition takes place as a consequence of GABA binding to its receptors — the so-called GABA_A receptors (Fig. 1). The result is a decrease in the probability that a neuron reaches its threshold for firing an action potential, the signal of a nerve cell. In an earlier study⁵, the same group described two distinct types of inhibition in a group of neurons — the cerebellar granule cells — that are involved in coordinating movements. 'Phasic' inhibition of these cells is mediated by discrete pulses of high concentrations of GABA released at synapses, where GABA acts on synaptic GABA_A receptors. In contrast, 'tonic' (continuous) inhibition is due to the persistent activation of extrasynaptic GABA_A receptors by the low concentrations of GABA in the extrasynaptic space — rather like the continuous, muffled music from many concert halls picked up by the microphones.

The GABA_A receptors underlying these two forms of inhibition differ in terms of their molecular composition and proper-

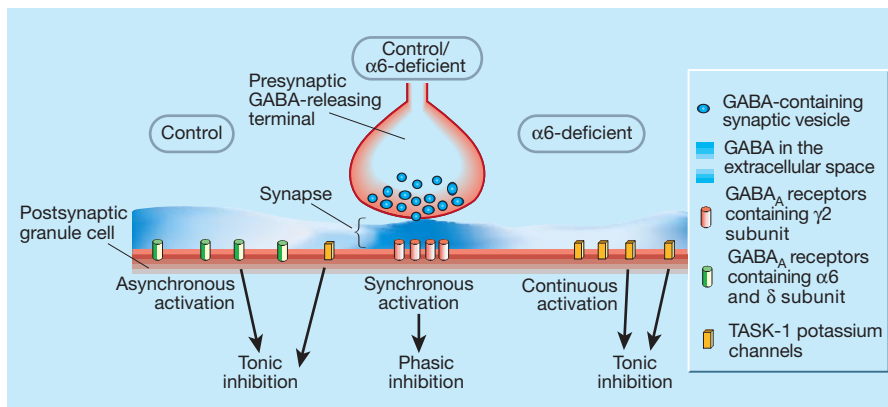


Figure 1 Background and synaptic inhibition of neuronal firing. The excitability of cerebellar granule neurons is regulated by inhibition, which is generated by the opening of certain ion channels. This increases the membrane's electrical 'leakiness' (conductance), making the cells less responsive to excitatory inputs. Inhibition can result from the binding of the neurotransmitter GABA to its receptors, and/or from the opening of potassium channels. Centre, in control mice and those that lack the $\alpha 6$ subunit of the GABA_A receptor, the release of GABA from synapses opens intrasynaptic, $\gamma 2$ -subunit-containing GABA_A receptors, generating a quickly rising, quickly decaying (phasic) inhibitory conductance. Left, in control mice, GABA diffusing away from several synapses activates extrasynaptic GABA_A receptors in an asynchronous manner, generating noisy, tonic inhibition. This is akin to music from several concert halls being picked up by a microphone outside the halls. Right, in $\alpha 6$ -deficient mice the extrasynaptic GABA_A receptors containing the $\alpha 6$ and δ subunits are missing. But, as shown by Brickley *et al.*⁴, tonic inhibition is restored because the granule cells increase their expression of TASK-1 potassium channels.

ties⁶. Receptors containing $\alpha 6$ and δ subunits are only present outside synapses, whereas receptors comprising the $\gamma 2$ subunit are concentrated inside synapses⁷. As a result, the extrasynaptic receptors have higher sensitivity (affinity) for GABA, and are not desensitized by the prolonged presence of this neurotransmitter. This makes them ideal for mediating tonic inhibition. So, to determine the importance of tonic inhibition in controlling neuronal excitability, Brickley *et al.*⁴ used mice that had been engineered to lack the $\alpha 6$ subunit⁸.

The authors could not detect any GABA_A-receptor-mediated tonic inhibition in brain slices from the genetically altered mice, an observation that fits well with the known role of the $\alpha 6$ -containing extrasynaptic receptors in generating tonic inhibition. In control animals, drugs that block the GABA_A receptors have a significant effect on the amount of excitation required to bring a granule cell to its firing threshold. These drugs had no such effect in the $\alpha 6$ -deficient animals⁴. As it is only the tonic form of GABA_A-receptor-mediated inhibition that is missing in these mice, this lack of an effect shows that this type of inhibition controls the excitability of granule neurons.

But if extrasynaptic GABA_A receptors were truly important, one might expect that deleting them would result in severe alterations in excitability within the neuronal circuits of the cerebellum. This brain region is important in coordinating movements, so specific defects in the behaviour of $\alpha 6$ -deficient mice would be predicted. But these mice do not show any measurable

behavioural problems⁸. Had the story ended here, sceptics might have mused that tonic inhibition is not named background inhibition for nothing. Fortunately, Brickley *et al.*

avoided this hasty conclusion, and go on to provide a surprising solution to the puzzle.

The authors show that the amount of excitation needed for granule cells to emit an action potential when the GABA_A receptors are not blocked by drugs is similar in control and $\alpha 6$ -deficient mice. They argue — on the basis of their electrophysiological and molecular data — that a compensation mechanism is at work in mice lacking the $\alpha 6$ subunit (Fig. 1). They suggest that this mechanism involves the increased expression of a continuously active type of potassium channel. Their results point to the so-called TASK-1 channels⁹, which also inhibit cellular excitability in a continuous (tonic) manner.

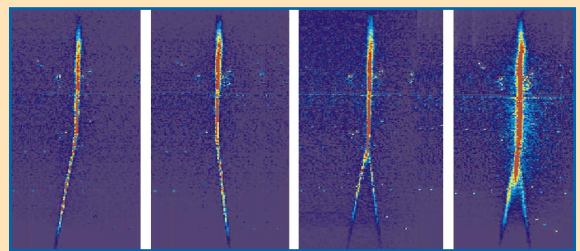
These data are the best evidence yet that the extrasynaptic GABA_A receptors regulate neuronal excitability. Given that biology tends not to fix things that are not important, the fact that neurons found a way to compensate precisely for the loss of extrasynaptic GABA_A receptors suggests just how significant the receptors are. Although Brickley *et al.* did their experiments in brain slices *in vitro*, compensation by the TASK-1 channels took place in the intact brain — a strong indication that tonic inhibition also existed *in vivo*. But the level of expression of many other GABA_A-receptor subunits (as well as that of the TASK-1 channels) is also altered in the $\alpha 6$ -deficient mice¹⁰.

Atom optics

A matter of choice

Similar to how light fibres transformed optics, physicists hope guided atom beams will revolutionize atom optics. Two groups have developed an atomic analogue of a beam splitter, which until now has only been used to separate atoms moving in free space. By guiding atoms along a surface, greater control has been achieved by D. Cassettari *et al.* (*Phys. Rev. Lett.* **85**, 5483–5487; 2000) and D. Müller *et al.* (<http://xxx.lanl.gov/abs/physics/0003091>).

Modern atom interferometers use beam splitters to split up matter waves and recombine them to produce sensitive interference patterns, used in precision measurements of gravity and the rotation of the Earth. But such free-space beam splitters can only separate beams by small angles thus limiting the scope of the experiments.



In their study, Cassettari *et al.* guided cold lithium atoms along a Y-shaped current-carrying wire patterned on the surface of a semiconductor chip. Fluorescence images taken of the wire show the beam splitter in action (see above). When current was passed through only one of the output wires, the atoms take the left or right fork (left two images). By varying the current in each arm, the splitting fraction of the atoms can be changed continuously (50:50 in right images). The angle between the two beams is 15°.

In similar work, Müller *et al.* split a beam of cold rubidium atoms along two current-carrying wires on a glass substrate. Neither of these studies split a coherent matter wave — in which all the atoms are in step with one another. This has been shown for atoms in free space, but needs to be done with guided atoms if they are to be used in precision atom interferometers. One solution might be to use a coherent atom source, such as a Bose–Einstein condensate.

Sarah Tomlin

Developing 'inducible-knockout' mice, in which the expression of the $\alpha 6$ or δ subunits can be stopped quickly, might avoid the problems of compensatory alterations and may clarify the roles of tonic inhibition in brain function.

The next major advance in this field might not come from studies of genetically altered mice. The development of drugs that specifically block receptors containing the δ subunit, for example, would provide tools for unravelling the precise function of background inhibition, not just in the cerebellum but also in other brain areas where such receptors are expressed extrasynaptically. The task of understanding the function and importance of extrasynaptic GABA_A receptors is an exciting one, and it is clear that neuroscientists have already lost their inhibition about listening to background noise. ■

Ivan Soltesz is in the Department of Anatomy and Neurobiology, University of California, Irvine, California 92697-1280, USA.

e-mail: isoltesz@uci.edu

Zoltan Nusser is in the Laboratory of Cellular Neurophysiology, Institute of Experimental Medicine, Hungarian Academy of Sciences, Szegony Street 43, 1083 Budapest, Hungary.

e-mail: nusser@koki.hu

1. Somogyi, P., Takagi, H., Richards, J. G. & Mohler, H. *J. Neurosci.* **9**, 2197–2209 (1989).
2. Soltesz, I. *et al.* *Eur. J. Neurosci.* **2**, 414–429 (1990).
3. Nusser, Z., Roberts, J. D. B., Baude, A., Richards, J. G. & Somogyi, P. *J. Neurosci.* **15**, 2948–2960 (1995).
4. Brickley, S. G., Revilla, V., Cull-Candy, S. G., Wisden, W. & Farrant, M. *Nature* **409**, 89–93 (2001).
5. Brickley, S. G., Cull-Candy, S. G. & Farrant, M. *J. Physiol. (Lond.)* **497**, 753–759 (1996).
6. Rossi, D. J. & Hamann, M. *Neuron* **20**, 783–795 (1998).
7. Nusser, Z., Sieghart, W. & Somogyi, P. *J. Neurosci.* **18**, 1693–1703 (1998).
8. Jones, A. *et al.* *J. Neurosci.* **17**, 1350–1362 (1997).
9. Duprat, F. *et al.* *EMBO J.* **16**, 5464–5471 (1997).
10. Nusser, Z. *et al.* *Eur. J. Neurosci.* **11**, 1685–1697 (1999).

Quantum engineering

Squeezing entanglement

Nick Bigelow

Quantum entanglement between two particles is a spooky connection that means measuring one has an instant effect on the other. Connecting many atoms in this way would be the first step towards a quantum computer.

If a street magician with two identical coins told you he could predict which way up your coin would land — heads or tails — simply by tossing his coin first, you probably wouldn't believe him. But what if he told you that, because of the laws of physics, your toss had to turn out the same as his toss? Not convinced, you try it and find that, yes, it is true. And it remains true, time after time, toss after toss. By some mechanism, there is a surprising correlation between the behaviour of the two coins. What's going on? Well, it could be that these two coins have somehow been prepared in a remarkable quantum state known as an entangled state. On page 63 of this issue Sørensen *et al.*¹ provide physicists with an exciting new recipe for creating such entangled states from an unusual sample of atoms known as a Bose–Einstein condensate.

The concept of entanglement is one of the most fundamental features of quantum mechanics, yet it is one of the most puzzling, non-intuitive and 'non-classical' aspects of the theory. The consequences of entanglement are so disturbing that Albert Einstein called them "spooky action-at-a-distance". But is entanglement real? Can we actually create entangled states? More importantly, can we observe the effects of entanglement? The answer to all of these questions is yes, at least on the rather remote and microscopic scale of a single pair of photons. More recently, entanglement has also been demonstrated

using an ensemble of four carefully prepared atoms^{2,3}. So far, though, entanglement has not been observed in any macroscopic (human-sized) system.

The physics of entangled states is also at the heart of a new generation of futuristic technologies, including recent plans for quantum computers and strategies for quantum teleportation. Making entanglement a tangible, exploitable phenomenon, however, requires the creation of entangled states of many particles — entanglement on a macroscopic scale. Moreover, it is important to achieve this with massive particles that can easily be stored and transported, rather than with photons, which have no mass. One of the exciting aspects of the work of Sørensen *et al.*¹ is that, by following their guidelines, researchers may soon be able to do just that — entangle the many particles within a Bose–Einstein condensate (BEC). A BEC is a large sample of particles (as many as 10 million ultracold atoms) that share exactly the same quantum state.

For the purposes of entanglement, theorists^{1,4–6} are especially interested in a type of BEC in which the atoms have multiple internal states. This was first achieved experimentally for a 'double condensate' composed of two clouds of rubidium atoms,⁷ each cloud having a different internal spin, which can be thought of as a tiny bar magnet. More recently, researchers have been able to transfer a BEC of sodium atoms from a magnetic trap

into an optical trap formed from a focused laser beam⁸. Unlike a magnetic trap, this new laser-trapping technique is not sensitive to the spin state of the atoms, so physicists can vary the complex magnetic properties of these 'spinor' condensates^{9,10}.

In the quest to create entangled states for large collections of atoms, the 'squeezed' state is of particular interest. To appreciate the relationship between squeezing and entanglement, it is important to have a sense for quantum 'noise'. At the heart of quantum theory is the idea that nature is inherently probabilistic. In classical physics you can predict the outcome of a coin toss if you know the exact starting conditions. But in quantum theory you can speak only of the probability of a certain outcome, no matter how much detail of the problem is known. Inherent in this picture is the idea that the measurable properties of a given state are accompanied by unavoidable fluctuations.

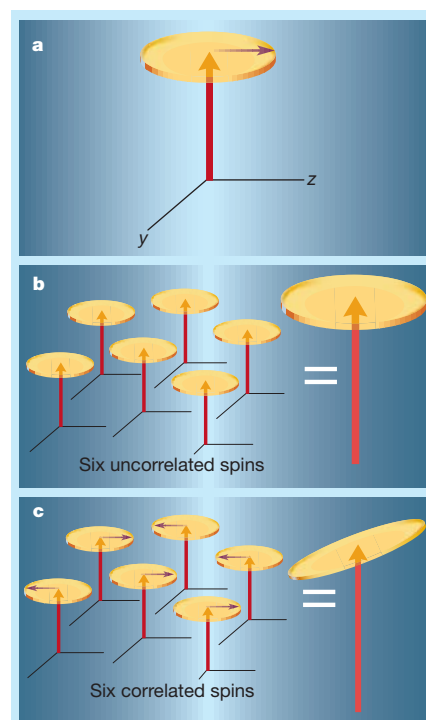


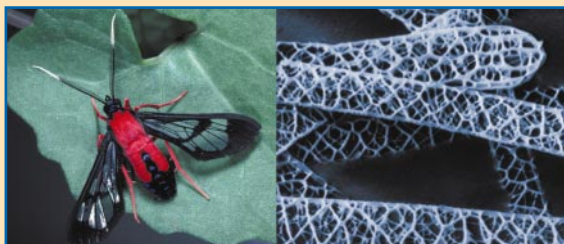
Figure 1 Putting the squeeze on spin. Simple particles, such as electrons, have a quantum mechanical 'spin', which can be either up or down. **a**, This quantum spin is aligned mostly in the 'up' direction, but there is a quantum mechanical uncertainty in the component of the spin in the transverse direction, which is represented by a small circular disk. **b**, In a dilute gas of N atoms with uncorrelated spins, the collective uncertainty is a disk of diameter $\sqrt{N}$ times each individual uncertainty disk. **c**, In a gas of atoms with correlated spins, such as the Bose–Einstein condensate modelled by Sørensen *et al.*¹, the uncertainty disk becomes an ellipse, which is narrower along the y -axis than the z -axis. This means that measuring the left–right component of the spin is 'noisier' than the in–out component — it has been squeezed.

Entomology

The alkaloid defence

The moth *Cosmosoma myrodora*, pictured left, is visually striking. And, as revealed by William E. Conner and colleagues in the *Proceedings of the National Academy of Sciences USA* (97, 14406–14411; 2000), its behaviour is also rather unusual. The male moths seem to have developed a chemical system to protect themselves, their female mates and their offspring from predation by spiders.

Conner *et al.* provide evidence that male — but few female — *C. myrodora* feed on the fluid secreted from certain plants (perhaps *Eupatorium capillifolium*). The alkaloid compounds thus ingested become particularly concentrated in a mass of filaments (pictured right) in the abdomen of the moths. The



authors assume that this protects the males from spiders such as *Nephila clavipes* — when the moths were fed a similar alkaloid in the lab, the spider cut the moths free from its web rather than eat them.

Before mating, male *C. myrodora* release some of their filaments, covering their chosen female. This seems to protect the female from spiders, too. The females may also receive alkaloids from the males' sperm, and in turn pass

on some of these protective chemicals to their eggs.

But questions remain. Do the females use receipt of alkaloids as a measure of a male's 'worth'? Females did seem to prefer males that had released filaments, but it is not certain if females could discern whether the filaments were laden with alkaloids. It is also not clear which plants the moths feed on in the wild, because the moths are rare and hard to spot. **Amanda Tromans**

shining a judiciously tailored microwave field on a BEC and letting atoms in the condensate collide with each other, it is possible to achieve entanglement-induced squeezing.

This route to entanglement not only demonstrates the sort of large-scale quantum engineering needed for quantum-information applications, but also has potentially important consequences in other areas, such as precision atomic clocks^{12,13}. The performance of sophisticated laser-cooled atomic clocks is already close to the limits set by quantum noise¹⁴, a limitation that could be overcome if a spin-squeezed atomic BEC is used to run the clock.

Although recent experiments in our group and elsewhere have shown that a BEC is not absolutely required to create a spin-squeezed atomic vapour^{15,16}, the idea of marrying the power of entanglement with the remarkable properties of a BEC offers outstanding possibilities for creating a new generation of non-classical atomic states^{1,4–6}. One day, we may even hear about entanglement of another macroscopic form of matter — the bulk sample of metal found in a simple pair of coins. ■

Nick Bigelow is in the Department of Physics and Astronomy and the Laboratory for Laser Energetics, PO Box 270171, University of Rochester, Rochester, New York 14627, USA.

e-mail: nbig@le.rochester.edu

1. Sørensen, A., Duan, L.-M., Cirac, J. I. & Zoller, P. *Nature* **409**, 63–66 (2001).
2. Sackett, C. A. *et al.* *Nature* **404**, 256–259 (2000).
3. Rauschenbeutel, A. *et al.* *Science* **288**, 2024–2028 (2000).
4. Pu, H. & Meystre, P. *Phys. Rev. Lett.* **85**, 3987–3990 (2000).
5. Law, C. K. *et al.* *Phys. Rev. A* (in the press).
6. Raghavan, S. *et al.* *Opt. Comm.* (in the press).
7. Myatt, C. J. *et al.* *Phys. Rev. Lett.* **78**, 1880–1883 (1997).
8. Stamper-Kurn, D. *et al.* *Phys. Rev. Lett.* **80**, 2097–2100 (1998).
9. Ho, T.-L. *Phys. Rev. Lett.* **81**, 742–745 (1998).
10. Ohmi, T. & Machida, K. *J. Phys. Soc. Jpn* **67**, 1822–1826 (1998).
11. Kitegawa, M. & Ueda, M. *Phys. Rev. A* **47**, 5138–5143 (1993).
12. Wineland, D. J. *et al.* *Phys. Rev. A* **50**, 67–88 (1994).
13. Boyer, P. & Kasevich, M. *Phys. Rev. A* **56**, R1083–R1086 (1997).
14. Santarelli, G. *et al.* *Phys. Rev. Lett.* **82**, 4619–4622 (1999).
15. Hald, J. *et al.* *Phys. Rev. Lett.* **83**, 1319–1322 (1999).
16. Kuzmich, A. *et al.* *Phys. Rev. Lett.* **85**, 1594–1597 (2000).

These fluctuations are an expression of the Heisenberg uncertainty principle, and they set a quantum noise limit on the accuracy of any precision measurement. In a squeezed state, this quantum noise is 'squeezed', or redistributed in the system, so that some measurable properties become 'quieter', whereas other properties become 'noisier'.

The states studied by Sørensen *et al.*¹ are 'spin squeezed'^{11,12}. In the quantum world we often represent an atomic spin by an arrow (Fig. 1a). For the simplest spins, like those of electrons, these arrows can point either up or down. Now, applying the uncertainty principle, we find that the transverse part of the spin (the part not exactly in the up–down direction) is uncertain by an amount represented by a small disc. In other words, if we try to find out whether the spin is angled left or right, or in or out of the page, we find that we cannot specify both the amount of left–rightness and in–outness at the same time. In the language of quantum noise, the transverse spin is 'noisy' in the left–right and in–out directions.

For a gas made up of many atoms, each with their own spin, the collective atomic spin is represented by one big arrow and one big uncertainty disc (Fig. 1b). A key idea exploited by Sørensen *et al.* is that if we entangle the individual atomic spin states, by introducing carefully tailored correlations between the individual atomic spins, then the collective spin state of the vapour can be squeezed. In Fig. 1c the transverse compo-

nents of the individual atomic spins preferentially add up in the left–right direction, as opposed to the in–out direction, changing the uncertainty disc of the total spin from a circle into an ellipse — it is now squeezed.

One consequence of this spin-squeezing is that the quantum noise involved in measuring in–outness can be made smaller than that for measuring left–rightness. This entangled squeezed state provides a way to break what is known as the standard quantum limit for the measurement of one component of the collective spin (the standard quantum limit is the diameter of the unsqueezed uncertainty circle in Fig. 1b). The essence of Sørensen *et al.*'s idea is that by

Mammalian evolution

Relationships to chew over

Anne Weil

Did advanced mammals evolve on the southern continents and then move north? Not according to a new study, which concludes that such mammals evolved in both the south and the north.

There are three groups of living mammals — placentals, marsupials and the monotremes. The first two, along with some mammalian fossil relatives, have so-called 'tribosphenic' teeth, which provide a highly efficient way of chopping and grinding food. Monotreme ancestors are also thought to have possessed such teeth.

On page 53 of this issue, Luo *et al.*¹ argue that mammals with tribosphenic teeth evolved not once but twice, after the super-continent of Pangaea pulled apart more than 160 million years ago. According to their hypothesis, one lineage radiated across the southern landmass of Gondwana but is represented today by only the platypus and

echidnas, which together constitute the monotremes. The other lineage, isolated on northern Laurasia, gave rise to living marsupial and placental mammals (Fig. 1a). Luo and colleagues' taxonomy is based on an extensive analysis of teeth and skeletal remains. But it is radically at odds with other interpretations of fossil data^{2,3}, and also with an evolutionary history reconstructed from sequences of mitochondrial DNA⁴ (Fig. 1b,c).

In the past three years, there have been reports^{2,3} of two significant discoveries of early Gondwanan mammals. In both cases, the fossil animals are surprisingly advanced, given their antiquity. Rich *et al.*² placed the 120-million-year-old *Ausktribosphenos* firmly within the Tribosphenida, a clade of mammals that includes marsupials and placentals, and that is diagnosed by the presence of tribosphenic teeth. Indeed, they suggested that *Ausktribosphenos* might be an early representative of the placental lineage. This seems arguable, however — *Ausktribosphenos* has peculiar teeth, and a jaw structure that is in some respects primitive^{2,5}.

The other fossil, *Ambondro*, known from a 167-million-year-old jaw from Madagascar, was described by Flynn *et al.*³. They placed *Ambondro* within the Tribosphenida, too, fuelling speculation that the break-up of Pangaea had not been much of a barrier to mammalian dispersal, or even that Tribosphenida originated in the south. In placing *Ausktribosphenos* and *Ambondro* in the Tribosphenida, Rich *et al.* and Flynn *et al.* upheld a long-standing view that the tribosphenic dentition evolved once. Their findings contradicted previous fossil evidence for a Laurasian origin of the Tribosphenida.

The extant platypus and echidnas are specialized for feeding on invertebrates and have no teeth. But a toothed monotreme, *Steropodon*, is known from a partial jaw, which was discovered in an Australian opal mine and dates to between 113 million and 97.5 million years ago⁶. Analyses that included *Steropodon* indicated that monotremes were more primitive than mammals with tribosphenic teeth; they have been considered to possess a precursor morphology to the more advanced condition. Now Luo *et al.*¹ have included *Ausktribosphenos*, *Ambondro* and *Steropodon* with other Mesozoic mammals in three phylogenetic analyses that include characters from the teeth and jaws, and also from the rest of the skeleton. Each analysis produces the same elegant result: the Gondwanan animals, including monotremes, are closely bound together in a clade that is widely separated from that of the marsupial and placental mammals. In this new model, mammals evolved tribosphenic teeth in parallel, on either side of a vast and widening gulf — the Tethys Sea that divided Gondwana and Laurasia.

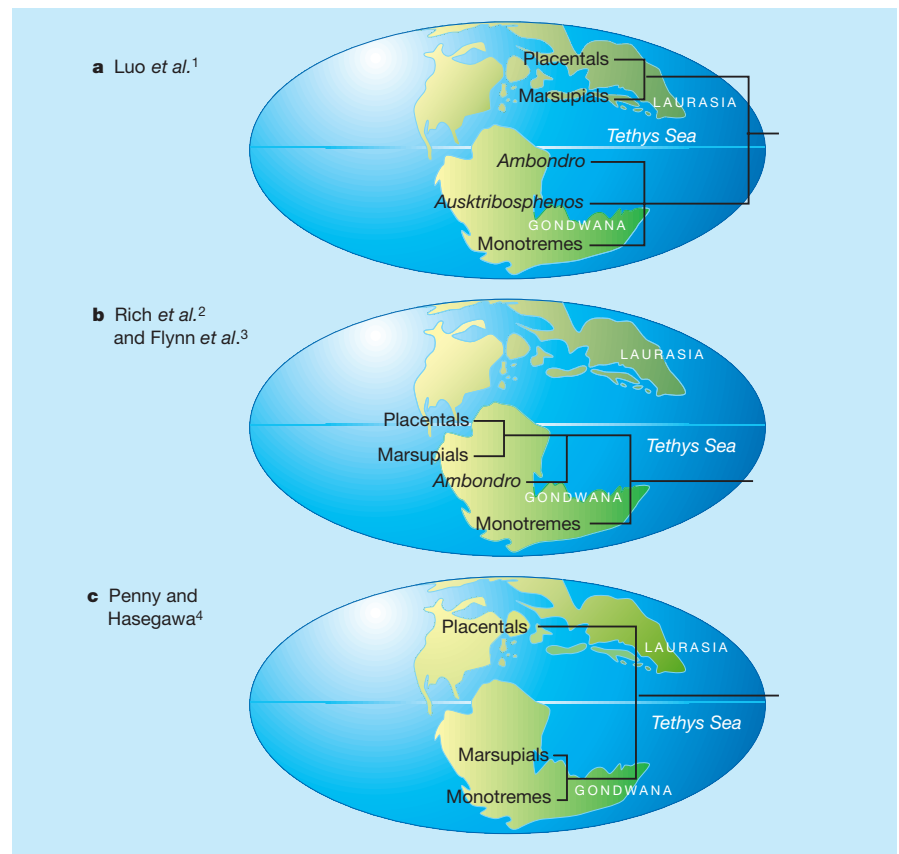


Figure 1 Proposed evolutionary relationships and origins of living mammals — placentals, marsupials and monotremes — and some fossil relatives. a, Luo *et al.*¹ retain a widely accepted theory of early mammalian biogeography, but propose that complex tribosphenic teeth evolved twice. b, This contrasts with a composite phylogenetic tree inferred from relationships favoured by Rich *et al.*² and Flynn *et al.*³, in which the fossil *Ausktribosphenos* is included among placentals. In this scheme, tribosphenidans would have dispersed northwards from a southern point of origin. Both a and b are based on fossil data. c, Both views conflict with that of Penny and Hasegawa⁴, who considered molecular data without discussing biogeographical or fossil evidence. The geographical distribution of origins shown in c is that favoured by Gregory for Marsupionta⁹.

This view of events does not require as unlikely a convergence as it might seem. Early mammals were small and endothermic (loosely speaking, warm-blooded), with high surface-to-volume ratios. They probably had high metabolic rates and correspondingly high nutritional requirements. Living shrews, which face the same constraints, have prodigious appetites. So survival probably depended on efficient food processing, and the tribosphenic dentition provides just that.

As a tribosphenic mammal bites down, a large cusp on the upper molar settles, mortar-like, into a pestle-like basin of the lower molar. Simultaneously, notched shearing crests on the sides of the triangular upper molar scissor against those of the lower molars. This combination of shearing and grinding has long been considered a key innovation in the clade containing marsupial and placental mammals — indeed, as the innovation that was possibly most significant in the spread and diversification of mammals. So advantageous is it that a third, entirely extinct, mammalian lineage, repre-

sented by the fossil *Shuotherium*, seems independently to have evolved a 'pseudo-tribosphenic' system that functioned identically, although the positions of the lower molar basin and shearing crests were reversed⁷.

Nor is the hypothesis that marsupials and placentals originated on the northern continents startling. Although most living marsupials are restricted to the Southern Hemisphere, the fossil record of the early marsupial lineage (Metatheria) is all Laurasian: the earliest metatherians are Asian, and the most diverse are North American. Marsupials are thought to have spread into Gondwana sometime between 84 and 64 million years ago, moving between Australia and South America by way of Antarctica⁸.

This view superseded an earlier hypothesis of mammalian relationships, espoused by W. K. Gregory before the acceptance of plate tectonics and before the discovery of toothed monotremes, *Ambondro*, *Ausktribosphenos* or many fossil metatherians. Gregory's hypothesis was that monotremes and marsupials are more closely related to each other than they are to placentals, forming a clade

dubbed Marsupionta⁹. Ironically, as new fossil evidence makes the Marsupionta hypothesis ever more obsolete, some⁴, although not all¹⁰, phylogenies based on gene sequencing appear to support it.

With three competing hypotheses on the table, we can only be sure that more genes will be sequenced and more fossils will be found. What Luo *et al.*¹ offer is a well-supported theoretical framework in which everything seems to make sense: unusual morphological features of the early southern mammals are explained in a phylogenetic context, and biogeographical considerations do not require mouse-sized animals to cross the Tethys Sea. This framework is certain to be useful, but it will also engender controversy. And, like Gregory's Marsupionta, it may eventually turn out to be wrong. The Meso-

zoic of Gondwana is not well known — surprises are no doubt still in store. ■

Anne Weil is in the Department of Biological Anthropology and Anatomy, Duke University, 08 Biological Sciences Building, Science Drive, Box 90383, Durham, North Carolina 27708-0383, USA. e-mail: anneu@duke.edu

1. Luo, Z., Cifelli, R. L. & Kielan-Jaworowska, Z. *Nature* **409**, 53–57 (2001).
2. Rich, T. H. *et al. Science* **278**, 1438–1442 (1997).
3. Flynn, J. J., Parrish, J. M., Rakotosamimanana, B., Simpson, W. F. & Wyss, A. E. *Nature* **401**, 57–60 (1999).
4. Penny, D. & Hasegawa, M. *Nature* **387**, 549–550 (1997).
5. Kielan-Jaworowska, Z., Cifelli, R. L. & Luo, Z. *Lethaia* **31**, 267–268 (1998).
6. Archer, M., Flannery, T. F., Ritchie, A. & Molnar, R. *Nature* **318**, 363–366 (1985).
7. Chow, M. & Rich, T. H. *Austr. Mammal.* **5**, 127–142 (1982).
8. Woodburne, M. O. & Case, J. A. *J. Mamm. Evol.* **3**, 121–161 (1996).
9. Gregory, W. K. *Bull. Am. Mus. Nat. Hist.* **88**, 1–52 (1947).
10. Killian, J. K. *et al. Mol. Cell* **5**, 707–716 (2000).

Immunology

It takes more than two to tango

Ronald H. Schwartz

In vivo studies of a pair of co-stimulatory molecules in the immune system of mice may further our understanding of allergic reactions and inflammatory immune responses in humans.

Most T cells of the immune system recognize antigens — substances that stimulate an immune response — as protein fragments. These fragments must be presented to T cells by molecules on the surface of another type of cell, an antigen-presenting cell (APC). But this recognition event alone is not enough to fully activate the T cell. Simultaneous 'co-stimulatory' signals from the APC are needed before the T cell can proliferate and become specialized to perform a particular function, such as

secreting intercellular signalling molecules called cytokines. On pages 97–109 of this issue^{1–3}, three groups investigate a receptor protein called ICOS, which occurs on T cells and recognizes a co-stimulatory signal on APCs. All three groups agree that experimentally inactivating the ICOS gene in mice profoundly impairs the animals' ability to produce certain types of cytokines and antibody molecules.

The first co-stimulatory pair of molecules to be identified were the receptor mol-

tein CD28 — expressed continuously on all mouse and many human T cells — and its two binding partners, CD80 and CD86. The expression of these binding partners, or ligands, is not continuous but rather can be induced on all APCs⁴ (Fig. 1). Together with signalling through the T cell's antigen receptors, this co-stimulatory set enhances the production of the cytokine interleukin-2, which encourages the growth and differentiation of T cells. Subsequent work uncovered a second molecule on the T-cell surface that recognizes CD80 and CD86. The expression of this molecule, called CTLA-4, is inducible and imposes negative feedback on the process of T-cell activation.

A completely different molecular pair was then also found to be essential for co-stimulation⁵. In this case, the co-stimulatory signal (CD40) is continuously expressed on the APC, and the expression of its receptor (CD154) can be induced on the T cell (Fig. 1). Signalling through this co-stimulatory pair creates a strong positive-feedback loop, augmenting the expression of CD80 and CD86 on the APC. Many other co-stimulatory molecules have since been described. One of the most recently discovered pairs — consisting of ICOS and its binding partner, B7RP-1 — is another positive regulatory set.

Why is co-stimulation so complex? Presumably, each molecular pair has some important, unique function and so needs to be regulated independently of the other sets. A major clue to the function of ICOS and its partner lies in where and how these molecules are expressed. ICOS is expressed on the T cell in response to signalling through CD28 and the T-cell antigen receptor⁶. More important, its binding partner has a unique expression pattern. B7RP-1 is found only on the types of APCs called B cells and macrophages, and not on the major category

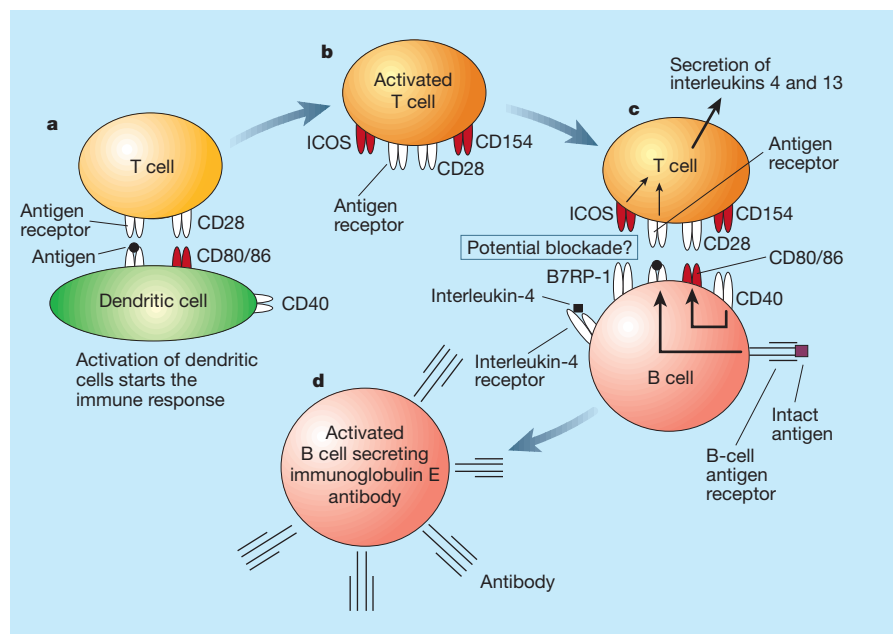


Figure 1 Co-stimulatory signals and immune responses. a, A dendritic cell (a type of antigen-presenting cell) is activated to express the co-stimulatory molecules CD80 and CD86. It can then present antigen to a T cell, which is activated by signalling through its antigen receptor and through CD28, the receptor for CD80/86. b, The activated T cell expresses all of its inducible co-stimulatory-signal receptors, such as ICOS. c, When the T cell next encounters a B cell that has become an antigen-presenting cell, the interactions of the co-stimulatory sets of molecules augment the production of cytokines (interleukins 4 and 13) by the T cell and of antibodies by the B cell (d). The interaction between ICOS and B7RP-1 is particularly potent in stimulating immunoglobulin E production, important in allergic reactions. Blocking this might be useful in treating asthma. Three new papers^{1–3} show that inactivating the ICOS gene in mice inhibits production of these cytokines and antibody, among other effects. Molecules in red are expressed only after cell activation.

of APC, a particular subset of dendritic cells⁷. Although other subsets of dendritic cells still need to be examined, these observations suggest that the ICOS–B7RP-1 pair is not involved in the initial activation of the T cell, but rather is important later on, when the B cells and macrophages are activated (Fig. 1).

The effects of inactivating ICOS in mice, described by Dong *et al.*¹, McAdam *et al.*² and Tafuri *et al.*³, are consistent with this idea. The 'knockout' mice have few germinal centres — the lymphoid factories where B cells manufacture antibody molecules. And certain types of antibodies, such as immunoglobulin E, are not generated at all in these mice. This can be partly explained by the decrease in production of specific cytokines (such as interleukins 4 and 13) by T cells. But the fact that there is decreased production of almost all antibody types suggests a more general defect.

One potential explanation is a slight decrease in the production of interleukin-2, described by Dong *et al.*¹ in *in vitro* analyses. But it might be more fruitful in the future to study the expression of chemokine receptors and their soluble ligands, chemokines. The chemokine-receptor family of cell-surface molecules, involved in the migration of immune cells during an immune response, is essential for the formation of germinal centres⁸. ICOS is expressed mainly on activated T cells that come into germinal centres, and its stimulation might amplify chemokine production to recruit more activated B and T cells to this site. In support of this idea, a 'transgenic' mouse expressing large amounts of soluble B7RP-1 has opposite characteristics to the knockout mouse — enlarged lymphoid tissue, with many germinal centres and antibody-forming cells⁷.

Alternatively, perhaps ICOS and its partner are mainly needed to augment antibody production by B cells: McAdam *et al.*² found that the knockout defects can be partially overcome by strongly activating the other co-stimulatory pathways. And soluble ICOS has no effects on the antiviral responses of cytotoxic T cells⁹. Effects on the activation of macrophages remain to be tested.

As well as their fundamental importance, these results^{1–3} have implications for the study of human diseases. The production of the immunoglobulin E antibody is an early feature of allergic responses to antigens such as pollen or dust mites. So the diminished production of this antibody in mice lacking ICOS would be expected to alleviate allergic conditions. Indeed, one recent study showed that administering a soluble form of ICOS — which acts as a decoy, preventing B7RP-1 from binding to ICOS on T cells — reduced the lung obstruction in an animal model of asthma¹⁰. This effect is best explained by another interesting feature of the expression pattern of B7RP-1. Inflammatory cytokines such as tumour-necrosis factor- α induce the

expression of B7RP-1 on fibroblast cells, found in the connective tissue of the lungs¹¹. Binding of this B7RP-1 to ICOS on nearby T cells might encourage these cells — when they recognize their antigen in the environment of the lung — to make cytokines such as interleukins 4 and 13. This would in turn encourage B cells to secrete immunoglobulin E. The soluble ICOS molecule would block this process.

On the opposite pole, cytokines such as interleukins 4 and 13 can also reduce inflammatory immune responses mediated by T cells. Indeed, when Dong *et al.*¹ inactivated the ICOS gene in mice immunized to show some of the symptoms of human multiple sclerosis, they found that this inflammatory disorder worsened. All in all, it seems that this co-stimulatory pair is well worth further

investigation in connection with several human immune-mediated diseases. ■

Ronald H. Schwartz is in the Laboratory of Cellular and Molecular Immunology, National Institutes of Health, National Institute of Allergy & Infectious Diseases, 9000 Rockville Pike, Bethesda, Maryland 20892-0420, USA.

e-mail: rs34r@nih.gov

1. Dong, C. *et al.* *Nature* **409**, 97–101 (2001).
2. McAdam, A. J. *et al.* *Nature* **409**, 102–105 (2001).
3. Tafuri, A. *et al.* *Nature* **409**, 105–109 (2001).
4. Abbas, A. K. & Sharpe, A. H. *Nature Med.* **5**, 1345–1346 (1999).
5. Grewal, I. S. & Flavell, R. A. *Annu. Rev. Immunol.* **16**, 111–135 (1998).
6. Hutloff, A. *et al.* *Nature* **397**, 263–266 (1999).
7. Yoshinaga, S. K. *et al.* *Nature* **402**, 827–832 (1999).
8. Cyster, J. *Science* **286**, 2098–2102 (1999).
9. Kopf, M. *et al.* *J. Exp. Med.* **192**, 53–61 (2000).
10. Coyle, A. J. *et al.* *Immunity* **13**, 95–105 (2000).
11. Swallow, M. M., Wallin, J. J. & Sha, W. C. *Immunity* **11**, 423–432 (1999).

Molecular electronics

Nanowires begin to shine

David H. Cobden

It is not easy finding a worthy successor to highly refined microchip technologies. But electronic devices built from molecular-scale components are fast becoming a good bet.

Unless you've been on a long vacation recently, you have probably heard a lot of hype about nanotechnology. Although many alternatives to existing microsystems are being considered, few would deny that there are formidable obstacles along this path. For example, the success of 'molecular electronics' — a potential successor to microelectronics — depends on the development of so-called bottom-up manufacturing techniques, in which molecular-scale components are chemically synthesized in large numbers and assembled into useful circuits¹. The difficulties are daunting but, as Duan *et al.* show on page 66 of this issue², real molecular-electronics devices are getting tantalizingly close. These authors have built nanometre-scale transistors and the smallest ever light-emitting diodes by exploiting two recent developments in the bottom-up approach to nanotechnology.

The first development was the perfection of techniques for growing nanometre-scale semiconductor wires. Until recently, such nanowires could be made only by depositing material on linear templates, such as natural fibres, the edges of thin films and the inside of membrane pores. The resulting wire structures are often polycrystalline and non-uniform. In contrast, the new methods require no template and produce nearly perfect single-crystal nanowires. The original idea goes back to the 1960s, when vapour-growth techniques were developed to produce crystalline semiconductors from hot gaseous reactants. It was discovered that

whiskers of the semiconductor would spontaneously grow out of gold particles³ placed in the reaction chamber. This is because atoms in the gases, such as gallium and arsenide, dissolve in the molten metal much more easily than they attach to the surface of existing crystals. Once the metal is saturated, the atoms readily pass out of solution onto the end of a monocrystalline whisker growing out of the particle (Fig. 1a).

The latest techniques are clever variations on this theme, which use much smaller (nanometre scale) metal particles that may be sitting on a substrate, suspended in a colloidal solution⁴, or floating in the hot gas. Duan and others in Charles Lieber's group at Harvard⁵ have perfected the latter method to produce a wide variety of semiconductor wires, with well-controlled diameter, length and composition. Lieber's group have even shown that chemical dopants (impurities that add or remove electrons) can be incorporated during the growth of the nanowires, thereby controlling whether the nanowires are n-doped (having extra conduction electrons) or p-doped (with some electrons removed to leave positively charged 'holes').

The second development concerns an old method for positioning objects such as nanowires where they are wanted. Because a nanowire is elongated, and so easily electrically polarized, it is attracted towards a high electric field, with which it lines up. So when a voltage is applied between two electrodes, a nearby nanowire suspended in liquid is drawn in to bridge the gap between them

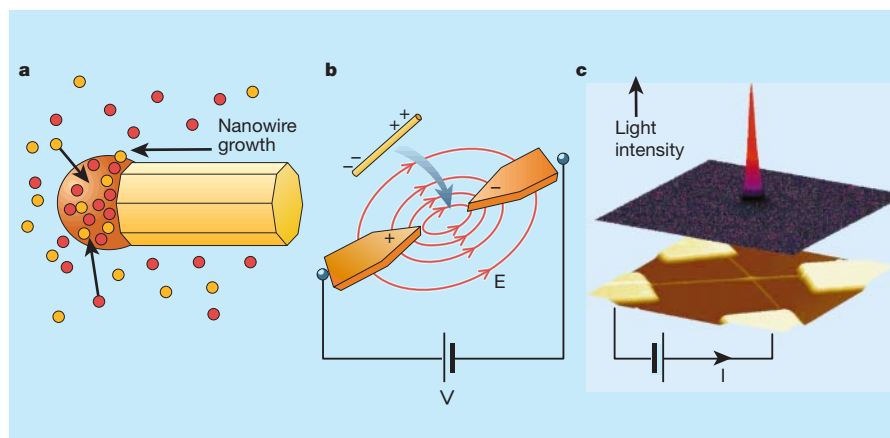


Figure 1 How to build the world's smallest light-emitting diode. **a**, A nanowire is grown from two reagents (red and yellow dots) with the help of a metal catalyst particle (orange). **b**, The nanowire is assembled between two metal electrodes using an electric field gradient. **c**, In the experiment of Duan *et al.*², a p-doped and an n-doped nanowire are crossed to form a nanoscale light-emitting diode. When a current is passed between them, electrons and holes are injected across the junction and recombine to emit light.

(Fig. 1b). Recent work has produced orderly rows of parallel single-nanowire bridges in this way⁶. Duan *et al.*² have now created a junction by placing a p-doped and an n-doped nanowire across each other (Fig. 1c).

Strong competitors for the same jobs as nanowires are single-walled carbon nanotubes, which are seamless hollow cylinders rather than solid rods. Both nanowires and nanotubes can be many micrometres long, making them far easier to work with than other popular molecular toys, such as conjugated molecules and nanocrystals, which are a thousand times shorter. This is why molecular field-effect transistors have previously been made only from nanotubes⁷, and now by Lieber's group from nanowires². Unlike nanowires, nanotubes have completely smooth surfaces free from dangling bonds. They can be either semiconducting or highly metallic, and also have tiny diameters (down to less than 1 nm), giving them fascinating one-dimensional electronic properties⁵. But control of the diameter, crystal-lattice orientation and doping seems to be much easier in nanowires, and their extra rigidity can make it easier to assemble them into position. Another big advantage is that semiconductors are known to emit light, something that has never been seen in carbon nanotubes.

Junctions between crossed metallic and p-doped semiconducting nanotubes have previously been shown to behave as electronic diodes⁷ — that is, current flows when a positive voltage is applied to the p-doped tube, but there is little response when the voltage is negative. The crossed nanowires of Duan *et al.* also make diodes, but with a remarkable new feature: when a positive voltage is applied to the p-doped nanowire, light is emitted from the junction. The luminescence process is just the same as in regular, large light-emitting diodes (LEDs), in

which electrons and holes injected across a p–n junction pair up and recombine to produce a photon. The authors also find that the frequency of the light increases if thinner nanowires are used, as the quantum states of the electrons inside them are squeezed into a smaller space, shifting their energy levels.

Assuming that their current-to-light efficiency can be improved — it is low because of

non-luminescent processes that radiate heat rather than light — the nano-LEDs created by Duan *et al.* offer a range of exciting new possibilities in optoelectronics. The colour (frequency) of light they emit can be varied by using nanowires of different thicknesses, as well as by the usual method of using semiconductors that naturally luminesce at different frequencies. If the non-luminescent processes can be strongly suppressed, a single nanowire may become the first one-dimensional laser. Also, these are the first LEDs to reach sub-wavelength dimensions — that is, they are smaller than the wavelength of the light they emit. Finally, if this assembly technique can be scaled up to produce a grid of crossed nanowires, they would form an extraordinarily high-resolution LED array. Such a device might have a chance of being the first bottom-up electronics ever to see the light.

David H. Cobden is in the Department of Physics, University of Warwick, Coventry CV4 7AL, UK.

e-mail: d.h.cobden@warwick.ac.uk

1. Joachim, C., Gimzewski, J. K. & Aviram, A. *Nature* **408**, 541–548 (2000).
2. Duan, X., Huang, Y., Cui, Y., Wang, J. & Lieber, C. M. *Nature* **409**, 66–69 (2001).
3. Wagner, R. S. & Ellis W. C. *Appl. Phys. Lett.* **4**, 89–90 (1964).
4. Holmes, J. D. *et al. Science* **287**, 1471–1473 (2000).
5. Lieber, C. M. *Solid State Comm.* **107**, 607–616 (1998).
6. Smith, P. A. *et al. Appl. Phys. Lett.* **77**, 1399–1401 (2000).
7. Fuhrer, M. *et al. Science* **288**, 494–497 (2000).
8. Dekker, C. *Physics Today* **52**, 22–28 (1999).

DNA replication

SOS polymerases

Fumio Hanaoka

The standard DNA-replicating machinery cannot copy damaged DNA, so SOS enzymes come to the rescue. It now seems, at least in some bacteria, that different enzymes are required for different types of damage.

Imagine that an injured person is carried into a hospital's emergency department. The doctor needs to assess the situation quickly, and to decide on the most appropriate treatment — stitches, a drug or maybe an operation? A similar problem occurs during DNA replication in living cells, when the replication machinery — consisting of enzymes called DNA polymerases — encounters a damaged DNA site. Standard DNA polymerases cannot continue to copy DNA once they reach a lesion. But every species has several polymerases that can do so, in a process called translesion synthesis¹. Why are there so many of these 'translesional' polymerases? And how does a cell decide which polymerase to use? Writing in *EMBO Journal* and *EMBO Reports*^{2,3}, Fuchs and colleagues reveal some of the answers for the bacterium *Escherichia coli*.

There are five DNA polymerases in *E. coli*. Three of these — DNA polymerases II, IV

and V — are induced as part of the cellular SOS response to DNA damage, and are involved in translesion synthesis. It was already known that DNA polymerase V (pol V)^{4,5} allows DNA replication to continue over bulky damage such as abasic lesions (when a base is lost from one strand of DNA) or errors produced by ultraviolet light⁶. Pol V is error-prone, and often leaves mutations in its wake within the newly synthesized strand opposite the damaged region. However, at certain types of lesion its activity is free from errors. It can also copy undamaged DNA; again, this is error-prone, although the mutations are 'untargeted' — they do not occur at a specific point in the DNA.

Pol IV, by contrast, is involved in getting replication started again after it stalls, and in inducing untargeted mutations when copying phage λ (ref. 7) — a small, circular DNA molecule (or plasmid) in *E. coli* that is replicated independently of the main chromo-

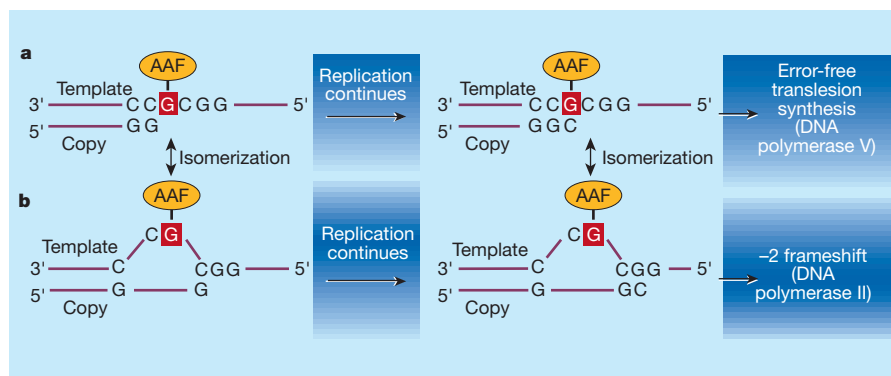


Figure 1 Translesional polymerases continue to replicate DNA even when it contains errors that make other DNA polymerases stall. There are several translesional polymerases in the bacterium *Escherichia coli*, and Napolitano *et al.*² show that the choice of polymerase depends on the nature of the DNA injury. One type of injury is that induced by the chemical *N*-2-acetylaminofluorene (AAF), which forms a bond with the fourth nucleotide of the sequence GCGGCC. a, DNA polymerase V is needed to copy this DNA in an error-free way. Incorporation of a cytosine residue (C) opposite the lesion by pol V generates a key replication intermediate. b, The template DNA and the copy in progress can slip relative to each other (producing a slipped intermediate), meaning that two of the bases in the template are not copied (a '– 2 frameshift'). The presence of the lesion thermodynamically favours the formation of a slipped intermediate (as shown by the occurrence of isomerization). DNA polymerase II is required in this pathway.

some. In fact, overexpression of pol IV results in a roughly 1,000-fold increase in mutations in plasmids⁸. Such mutations are likely to be useful in enabling a population to evolve quickly in response to environmental stress. Like pol V, pol IV is a bona fide DNA polymerase⁹. Little is known about the cellular function of *E. coli* pol II, however, even though it was discovered as long ago as 1970.

This is where Napolitano *et al.*² come in. They have looked at the frequency of mutations in various strains of *E. coli* in response to different types of chemically induced damage. They confirm that pol V is capable of error-free, as well as error-prone (mutagenic), translesion synthesis. But the interesting thing is that so are pol II and pol IV, and that the choice of polymerase depends on the nature of the DNA injury and the sequence in which it is found.

For example, the chemical *N*-2-acetylaminofluorene binds to the fourth nucleotide in the *NarI* DNA sequence (GCGGCC). It increases the rate of frameshift mutations here — in effect, the middle two bases of this sequence are not copied during DNA replication. Napolitano *et al.* show that pol II replicates DNA to which this chemical is bound, and induces the frameshift. Pol V, however, can replicate the GCGGCC sequence correctly, without introducing mutations (Fig. 1).

Similarly, both pol IV and pol V are required for error-free and mutagenic translesion synthesis when benzo(a)pyrene — a major cancer-causing component of cigarette smoke and car exhaust fumes — is used. This chemical joins to the third guanine of the sequence GGG. When *N*-2-acetylaminofluorene binds to a nucleotide in the

same stretch of DNA, only pol V is required for replication, and it induces a frameshift mutation. Here, perhaps the structural distortion of the DNA template by two different lesions in the same sequence requires a different combination of polymerases to achieve translesion synthesis. In any case, it seems that *E. coli* chooses from a pool of these polymerases to bypass different DNA lesions.

How is the relevant translesional polymerase targeted to a DNA lesion? Last year, Tang *et al.*¹⁰ showed that the SOS protein RecA activates pol V while directing it to a damaged DNA site. Wagner *et al.*³ now have biochemical data that provide an answer for pol IV. This *E. coli* enzyme is strictly distributive *in vitro* — it can copy only one base each time it binds to the DNA template. The authors could not detect a stable *in vitro* complex containing pol IV and DNA. The implication is that a co-factor helps pol IV to be targeted to its DNA substrate.

Wagner *et al.*³ proposed that the β -subunit of pol III — the main enzyme that replicates undamaged DNA — might be such a targeting factor. This subunit acts as a clamp to connect pol III to DNA. It now seems that this β -subunit does the same for pol IV, too³. When Wagner *et al.* added this clamp, together with the γ -subunit of pol III as a clamp 'loader', to an *in vitro* DNA-replication reaction, pol IV and DNA formed a stable complex. Pol IV could then incorporate 300–400 nucleotides per DNA-binding event. This value contrasts with that of 6–7 nucleotides obtained previously for pol IV in the presence of the β - and γ -subunits¹⁰, but it is not clear why these authors obtained such different results. In any event, it seems that pol IV and pol V

Daedalus

David Jones

David Jones, author of the Daedalus column, is indisposed.

require the β/γ complex of pol III and the RecA protein, respectively, to be recruited to their substrate DNA. Pol V then also requires the β/γ complex to carry out translesion synthesis.

Translesion synthesis is less accurate than normal DNA replication, but organisms probably chose it over the more dangerous state of leaving an unreplicated portion of DNA in a single-stranded form. However, translesion synthesis might have evolutionary benefits. As predicted more than a quarter of a century ago¹¹, the error-prone polymerases allow individual cells to mutate when their survival is threatened, increasing the genetic diversity and adaptability of an imperilled population.

On the other hand, one would not expect DNA polymerase η in humans to be as error-prone as its bacterial counterparts. This polymerase is important in preventing skin cancer — people suffering from xeroderma pigmentosum have a predisposition to skin cancers, and the gene mutated in these people is that encoding DNA polymerase η ^{12–14}. What is the difference between this enzyme and the *E. coli* translesional polymerases? To find out, we may have to wait until the structures of these enzymes in complex with their substrates have been determined. In the meantime, the work of Fuchs and colleagues^{2,3}, as well as that of Tang *et al.*¹⁰, provides much-needed information about how cells cope with DNA damage.

Fumio Hanaoka is at the Institute for Molecular and Cellular Biology, Osaka University, 1-3 Yamada-oka, Suita, Osaka 565-0871, Japan.
e-mail: fhanaoka@imcb.osaka-u.ac.jp

1. Friedberg, E. C., Feaver, W. J. & Gerlach, V. L. *Proc. Natl Acad. Sci. USA* **97**, 5681–5683 (2000).
2. Napolitano, R., Janel-Bintz, R., Wagner, J. & Fuchs, R. P. P. *EMBO J.* **19**, 6259–6265 (2000).
3. Wagner, J., Fujii, S., Gruz, P., Nohmi, T. & Fuchs, R. P. P. *EMBO Rep.* **1**, 484–488 (2000).
4. Tang, M. *et al. Proc. Natl Acad. Sci. USA* **96**, 8919–8924 (1999).
5. Reuven, N. B., Arad, G., Maor-Shoshani, A. & Livneh, Z. *J. Biol. Chem.* **274**, 31763–31766 (1999).
6. Friedberg, E. C., Walker, G. C. & Siede, W. *DNA Repair and Mutagenesis* (ASM, Washington DC, 1995).
7. Brotcorne-Lannoy, A. & Maenhaut-Michel, G. *Proc. Natl Acad. Sci. USA* **83**, 3904–3908 (1986).
8. Kim, S.-R. *et al. Proc. Natl Acad. Sci. USA* **94**, 13792–13797 (1997).
9. Wagner, J. *et al. Mol. Cell* **4**, 281–287 (1999).
10. Tang, M. *et al. Nature* **404**, 1014–1018 (2000).
11. Radman, M. in *Molecular and Environmental Aspects of Mutagenesis* (eds Sherman, F., Miller, M., Lawrence, C. & Tabor, W. H.) 128–142 (Thomas, Springfield, IL, 1974).
12. Masutani, C. *et al. EMBO J.* **18**, 3491–3501 (1999).
13. Masutani, C. *et al. Nature* **399**, 700–704 (1999).
14. Johnson, R. E., Kondratieva, C. M., Prakash, S. & Prakash, L. *Science* **285**, 263–265 (1999).

Tarantula peptide inhibits atrial fibrillation

A peptide from spider venom can prevent the heartbeat from losing its rhythm.

Atrial fibrillation is the most common sustained cardiac arrhythmia to occur in humans, secondary to valve disease, hypertension or heart failure. It is often associated with passive stretching of the atrial chamber arising from haemodynamic or mechanical dysfunction of the heart. Here we show that atrial fibrillation potentiated by dilatation in rabbit heart can be inhibited by blocking stretch-activated ion channels with a specific peptide from tarantula venom, without altering the resting action potential. Our findings open a window on cardiac arrhythmogenesis and point the way towards developing a new class of drugs.

Stretch-activated ion channels (SACs) are ubiquitous and have been studied in a variety of cardiac tissues^{1–3}. Non-selective cationic SACs are permeant to Ca^{2+} ions as well as to Na^+ and K^+ , whereas others are selective for K^+ and possibly Cl^- ions^{1,2}. According to computer models, the excitatory currents carried by SACs can generate fast arrhythmias^{4,5}, but a lack of specific inhibitors has prevented the acquisition of matching data.

GsMtx-4 is a small (relative molecular mass, 4K) peptide found in the venom of the tarantula *Grammostola spatulata* (Fig. 1a). It specifically blocks cationic SACs in astrocytes and inhibits volume-activated currents in both adult astrocytes and cardiocytes. As a member of the neuro-inhibitory 'cysteine-knot' family, GsMtx-4 has a binding affinity of about 500 nM for SACs in astrocytes. The specificity of its action is indicated by its lack of effect on the resting properties of rabbit ventricular cells and rat astrocytes⁶. At 4 μM (about 20 times the dose applied to suppress fibrillation), GsMtx-4 had no measurable effect on the action potential of isolated atrial heart cells (C. Baumgarten and H. Clemp, personal communication).

Using perfused rabbit hearts, we initiated fibrillation with a burst of high-frequency stimulation (Fig. 1b) at different right-atrial pressures^{7,8}. Stretching the chamber increased the incidence and duration of fibrillation (Fig. 1b–d). At pressures above 12 cm H_2O , the probability of sustained fibrillation (for longer than 60 s) approached unity. Perfusion with 170 nM GsMtx-4 suppressed both the incidence and duration of fibrillation in all hearts ($n = 10$). At pressures below 17.5 cm H_2O , sustained fibrillation was completely inhibited in all preparations (data not shown).

The gadolinium ion (Gd^{3+}), a common but non-selective SAC inhibitor, can also suppress stretch-induced cardiac fibrillation⁸, but its lack of specificity and non-applicability

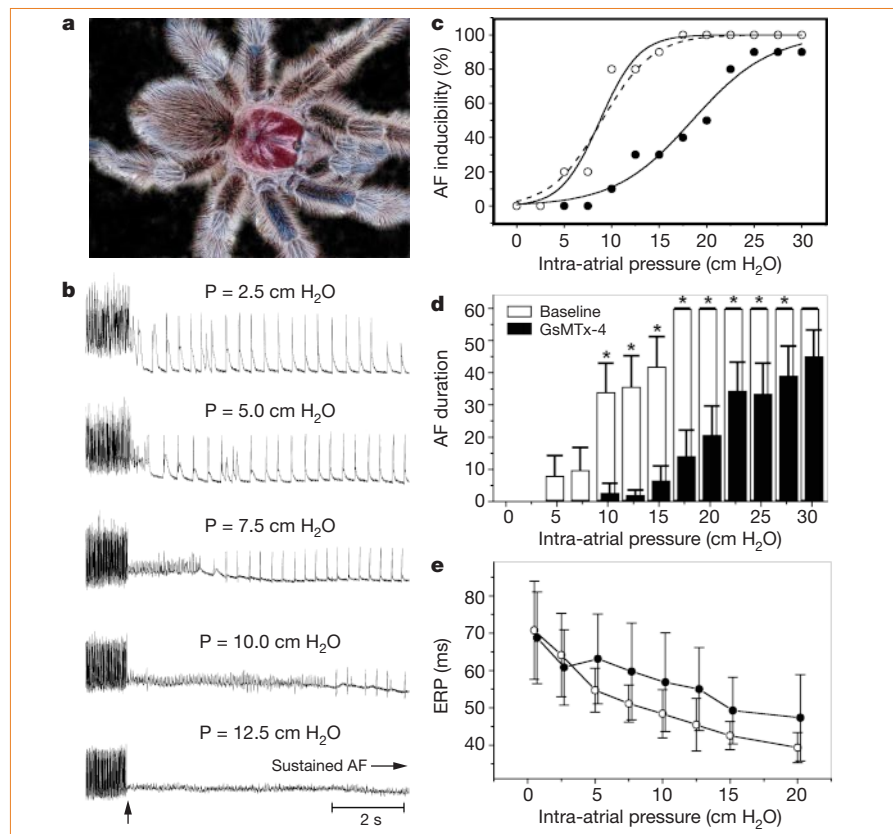


Figure 1 Inhibition of atrial fibrillation by GsMtx-4 during stretch. **a**, The spider *Grammostola spatulata*, whose venom is the source of the inhibitory peptide. **b**, Bipolar atrial electrograms showing an increase in atrial fibrillation (AF) with pressure, becoming sustained at 12.5 cm H_2O . The probability of inducing AF was increased by stimulating the heart with a short burst of high-frequency pacing before each measurement (arrow). **c**, Induction of AF lasting more than 2 s: circles, control; filled circles, in the presence of 170 nM GsMtx-4. Dashed line indicates the response after 20-min washout. **d**, Duration of AF ($n = 7$) as a function of pressure (mean $\pm$ s.e.). GsMtx-4 (170 nM) decreased the average time to spontaneous recovery from AF (asterisks, $P < 0.05$). **e**, GsMtx-4 did not block stretch-induced shortening of the refractory period ($n = 10$). For details of recording methodology, see ref. 8.

under physiological conditions⁹ restrict its testing on SACs. Despite their different chemical structures, however, both GsMtx-4 and Gd^{3+} suppress fibrillation in a similar manner without altering the stretch-dependence of the effective refractory period⁸ (Fig. 1e), perhaps because both reagents have a high positive-charge density.

The antifibrillatory effect we observe here cannot be driven by a change in surface-charge density, however, as this would alter the shape of the action potential. A different mechanism is also indicated by the persistence of the stretch-induced shortening of the refractory period while fibrillation is inhibited. It is possible that K^+ -selective SACs¹⁰, which are resistant to Gd^{3+} (ref. 11) and perhaps to GsMtx-4 as well, act to shorten the action potential under stretch.

The fundamental mechanism of fibrillation is viewed as the fragmentation of excitatory wavefronts, but the role of SACs in this process has not been studied directly.

SACs convert gradients of stress into gradients of electrical activity, and sustained mechanical gradients can develop, particularly during repolarization, when the systolic pressure and the wall tension are high while contraction is waning. These gradients are exacerbated in regions of the heart weakened by myocardial infarction, for example.

Stretch sensitivity is not unique to the atrium, so GsMtx-4 should act similarly on all chambers¹². As an investigative tool, this peptide should be useful for studying mechanical transduction at the molecular and whole-organism levels. We believe that GsMtx-4 could be the first of a new class of antiarrhythmic agents to be directed against the causes rather than the symptoms of fibrillation.

Frank Bode*†, Frederick Sachs‡, Michael R. Franz*

*Department of Pharmacology, Georgetown University, 3900 Reservoir Road, Washington DC 20007, USA

‡Department of Physiology and Biophysics,
State University of New York, 124 Sherman Hall,
Buffalo, New York 14214, USA

e-mail: Sachs@buffalo.edu

†Present address: Medizinische Klinik II,
Universitätsklinikum Luebeck,
Ratzeburger Allee 160, 23538 Luebeck, Germany

1. Ruknudin, A., Sachs, F. & Bustamante, J. O. *Am. J. Physiol.* **264**, H960–H972 (1993).
2. Hagiwara, N., Masuda, H., Shoda, M. & Irisawa, H. *J. Physiol. (Lond.)* **456**, 285–302 (1992).
3. Sato, R. & Koumii, S. *J. Membr. Biol.* **163**, 67–76 (1998).

Photonic engineering

Aphrodite's iridescence

The most intense colours displayed in nature result from either multilayer reflectors or linear diffraction gratings^{1–3}. Here we investigate the spectacular iridescence of a spine (notoseta) from the sea mouse *Aphrodita* sp. (Polychaeta: Aphroditidae). The spine normally appears to be deep red in colour, but when light is incident perpendicular to the axis of the spine, different colours are seen as stripes running parallel to the axis of the spine; over a range of smaller incident angles, the complete visible spectrum is reflected with a reflectivity of 100% to the human eye. The simple structure responsible for this effect is a remarkable example of photonic engineering by a living organism.

Photonics is a branch of optics that concerns the control of photons, much as electronics depends on electrons, and photonic structures have been described in butterfly wings⁴. The sea mouse has iridescent spines situated at the base of the dorsal surface, where they can be viewed laterally, with less strongly coloured threads being seen more readily from above (see <http://www.physics.usyd.edu.au/~nicolae/submit.html>). The biological function of the iridescence is unknown, but it may be related to species recognition or to courtship.

We prepared a spine with strong reflectance in the red for electron microscopy by mounting it in resin and slicing it with a microtome into sections transverse to the long axis. These sections were of constant thickness and were mounted on grids for transmission microscopy.

The micrograph in Fig. 1a shows an array of hollow cylinders, with the long axis of the cylinders along the spine. The image is of the wall region of the spine, the centre of the spine being hollow. The hollow cylinders have thicker walls near the edge of the spine, possibly for mechanical rigidity, but after a few layers their wall thickness decreases to a constant value. The cylinders are close-packed, with each cylinder having six nearest neighbours. The spacing of adjacent layers of cylinders is 0.51 μm , and remains constant

4. Sachs, F. in *Cell Mechanics and Cellular Engineering* (eds Mow, V. C. et al.) 308–328 (Springer, New York, 1994).
5. Zabel, M., Koller, B. S., Sachs, F. & Franz, M. R. *Cardiovasc. Res.* **32**, 120–130 (1996).
6. Suchyna, T. M. et al. *J. Gen. Physiol.* **115**, 583–598 (2000).
7. Ravelli, F. & Allesie, M. *Circulation* **96**, 1686–1695 (1997).
8. Bode, F., Katchman, A., Woosley, R. L. & Franz, M. R. *Circulation* **101**, 2200–2205 (2000).
9. Caldwell, R. A., Clemon, H. F. & Baumgarten, C. M. *Am. J. Physiol.* **275**, C619–C621 (1998).
10. Small, D. L. & Morris, C. E. *Br. J. Pharmacol.* **114**, 180–186 (1995).
11. Kim, D. J. *Gen. Physiol.* **100**, 1021–1040 (1992).
12. Nazir, S. A. et al. *J. Physiol. (Lond.)* **494P**, 111–112 (1996).

throughout the wall cross-section, which contains 88 layers of close-packed voids. The cylinder walls are made of pure α -chitin^{5,6} (refractive index, 1.54; ref. 2).

To calculate the expected optical properties of the spine, we used a formulation for layered stacks of cylinder gratings⁷. Figure 1c shows the reflectance of a 500-layer stack of gratings with period $d=0.51 \mu\text{m}$, for radiation incident in the plane of Fig. 1a,

with its electric vector perpendicular to the plane ($E_{\perp}$). The grating consists of hexagonally packed cylindrical holes of radius $a=0.2 \mu\text{m}$, filled with sea water (refractive index, 1.33), in a matrix of chitin. The bulk reflectance of chitin immersed in sea water is 0.54%, so the strong reflectance evident in Fig. 1c can only be achieved through the coherent stacking of scores of layers to form the crystal shown in Fig. 1a. Were the layers to be of irregular separation or composition, the peak reflectance for a given number of layers would be much lower⁷ and the iridescence less pronounced.

We calculated the total reflectance of this structure, for normal incidence, for free-space wavelengths in the visible spectrum (from 0.4 to 0.7 μm). Figure 1c shows a strong maximum in reflectance around the wavelength $\lambda=0.633 \mu\text{m}$. To investigate the origin of this narrow-band reflectance, we calculated the photonic band diagram, which gives the frequencies

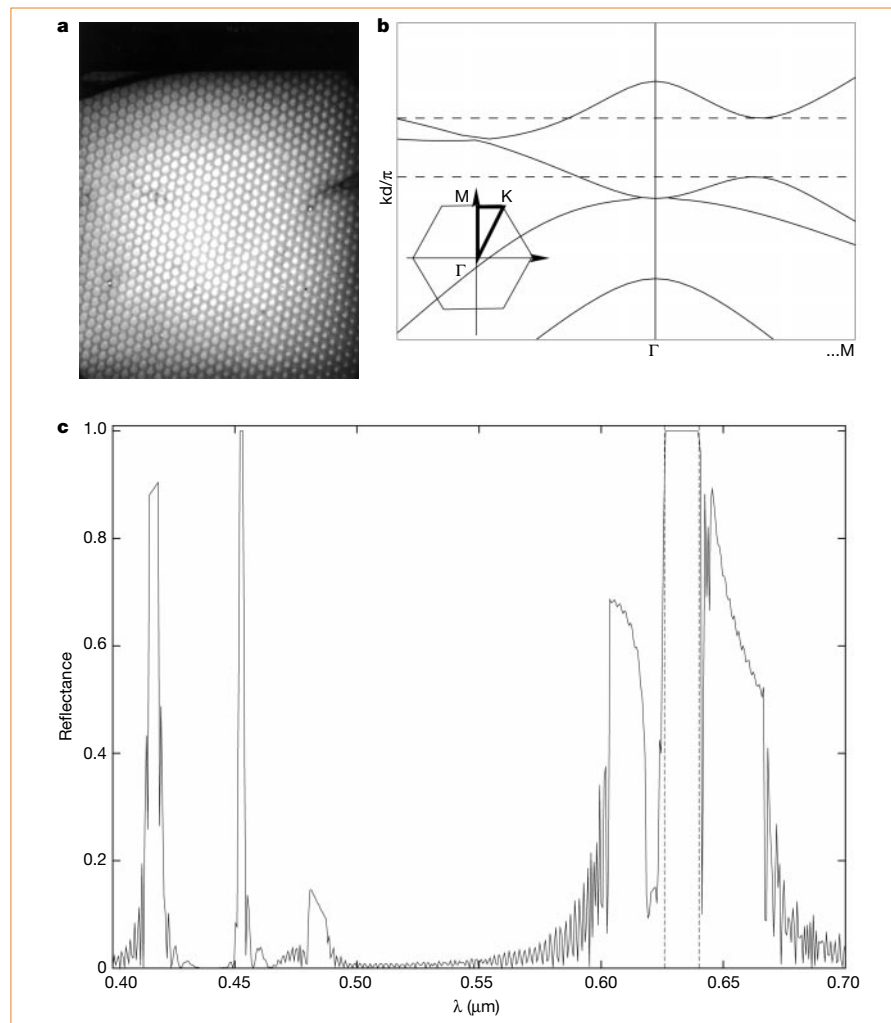


Figure 1 Structure and optical properties of a sea mouse hair from Palm Beach, New South Wales, Australia. **a**, Electron micrograph of a section of the spine (the dark regions are chitin). **b**, $E_{\perp}$ polarization photonic band structure for a hexagonal crystal of voids in chitin, near the four-fold degenerate point on the Γ -axis (dashed lines indicate a partial band gap). Γ , M and K are the corners of the irreducible part of the first Brillouin zone; k is the free space wavenumber. **c**, $E_{\perp}$ polarization total reflectance for normally incident radiation for a hexagonal structure having 500 layers of voids in chitin (dashed lines indicate the partial bandgap of **b**). Reflectance measurements using a Zeiss MMS spectrometer confirm a 100% reflectance in the red.

of light waves able to propagate freely in the structure, for the hexagonal array of voids (presumably water-filled) in a chitin matrix. This is shown in Fig. 1b for $E_{||}$ polarization, indicating that the region of enhanced reflectance in the red for the spine corresponds exactly to a partial photonic bandgap of the hexagonal structure (in the range Γ –M) — a circumstance in which light cannot propagate in a narrow range of wavelengths in the structure. Note that for $H_{||}$ polarization of incident radiation with the magnetic vector perpendicular to the plane of Fig. 1a, reflectance is similarly enhanced in the red, although there is not a clear-cut partial bandgap.

We have seen that the sea mouse achieves brilliant narrow-band coloration of its spines through a remarkable piece of photonic engineering. The regularity of the structure shown in Fig. 1a and the strong narrow-band reflectance shown in Fig. 1c suggest that growing optical filters by molecular self-assembly is a technological goal worth pursuing. These structures may have application in photonic communications, where there is much interest in fabricating photonic crystal fibres⁸ with similar morphology to that shown in Fig. 1a in order to improve bandwidth and nonlinear properties.

Andrew R. Parker*, Ross C. McPhedran†, David R. McKenzie‡, Lindsay C. Botten‡, Nicolae-Alexandru P. Nicorovici†

*Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

†School of Physics, University of Sydney, New South Wales 2006, Australia
e-mail: ross@physics.usyd.edu.au

‡School of Mathematical Sciences, University of Technology, Sydney, New South Wales 2007, Australia

1. Parker, A. R. in *Functional Morphology of the Invertebrate Skeleton* (ed. Savazzi, E.) 65–90 (Wiley, Chichester, 1999).
2. Land, M. F. *Progr. Biophys. Mol. Biol.* **24**, 75–106 (1972).
3. Parker, A. R. *Proc. R. Soc. Lond. B* **262**, 349–355 (1995).
4. Vukusic, P. *et al. Proc. R. Soc. Lond. B* **266**, 1403–1411 (1999).
5. Rouse, G. in *Invertebrate Zoology* (ed. Anderson, D. T.) 174–203 (Oxford Univ. Press, Melbourne, 1998).
6. Atkins, E. D. T., Dlugosz, J. & Foord, S. *Int. J. Biol. Macromol.* **1**, 29–32 (1979).
7. McPhedran, R. C. *et al. Aust. J. Phys.* **52**, 791–809 (1999).
8. Knight, J. C. *et al. Science* **282**, 1476–1478 (1998).

Oceanography

Vertical mixing in the ocean

The thermohaline circulation of the ocean results primarily from downwelling at sites in the Nordic and Labrador Seas and upwelling throughout the rest of the ocean. The latter is often described as being due to breaking internal waves. Here we reconcile the difference between theoretical and observed estimates of vertical mixing in the deep ocean by presenting a revised view of the thermohaline

circulation, which allows for additional upwelling in the Southern Ocean and the separation of the North Atlantic Deep Water cell from the Antarctic Bottom Water cell. The changes also mean that much less wind and tidal energy needs to be dissipated in the deep ocean than was originally thought.

Previous calculations of vertical mixing based on the stratification of the deep ocean^{1,2} assumed that a flux of 25 or 30 sverdrups (Sv) of water, made up of both deep and bottom waters, is injected at depths of about 4,000 metres and mixed upwards to depths near 1,000 m by turbulent mixing. Both reports conclude that the spatially averaged diapycnal (cross-density surface) mixing coefficient is $10^{-4} \text{ m}^2 \text{ s}^{-1}$.

However, observations of turbulence³ and dye diffusion⁴ in the deep ocean indicate that there exists a background diapycnal diffusivity of only $10^{-5} \text{ m}^2 \text{ s}^{-1}$, although much larger values are found in localized regions near rough topography⁵. The background value is consistent with mixing due to the background internal wave field, and the larger values are consistent with extra internal waves due to the interaction of currents with topography.

But it is not obvious that the latter is enough to raise diffusivity by an order of magnitude when averaged over the whole ocean. The extra power required to do this is also large⁶. If the efficiency is 20%, which is normally considered a maximum for the final stage of breaking internal waves, then the power required is 2.1 terawatts. This is just possible, given current estimates of the energy input from the wind and tides, but this figure does not allow for losses at other stages in the conversion process.

A contrasting view of the thermohaline circulation has come from low-resolution⁷ and high-resolution⁸ computer model studies of the ocean circulation. These show that between 9 and 12 Sv of deep water is brought to the surface by Ekman suction in the Southern Ocean. This is driven northwards in the surface Ekman layer and is reduced in density primarily by surface freshening. The model results also emphasize earlier observations⁹ that in the primary regions of bottom-water formation around Antarctica, the near-surface water masses have the same density as North Atlantic deep water. It is therefore not necessary for the bottom water to be mixed through the whole depth of the water column, only up to the level of the deep waters.

Using this new view of the thermohaline circulation, we need only consider the vertical mixing of the main deep-water mass, North Atlantic Deep Water, whose flux is estimated to lie between 14 and 17 Sv (ref. 10). Taking the larger of these two values and the smaller of the two model-based estimates of upwelling leaves a maximum of 8 Sv to be mixed vertically within the ocean.

The $10^{-5} \text{ m}^2 \text{ s}^{-1}$ background term can upwell 3 Sv, leaving 5 Sv to be upwelled by localized regions of intense mixing. If this view is correct, then the vertical mixing coefficient, averaged over the whole ocean, is less than $3 \times 10^{-5} \text{ m}^2 \text{ s}^{-1}$ and, assuming 20% efficiency, the total amount of extra energy required is less than 0.6 terawatts.

The revised values are consistent with existing observations of mixing within the ocean. They also emphasize again the importance of the Southern Ocean and imply that although further research is needed on the localized mixing in the deep ocean, such mixing does not control the thermohaline circulation.

D. J. Webb*, N. Sugimoto†

*Southampton Oceanography Centre, Empress Dock, Southampton SO14 3ZH, UK
e-mail: david.webb@soc.soton.ac.uk

†Center for Climate System Research, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan

1. Munk, W. H. *Deep-Sea Res.* **13**, 707–730 (1966).
2. Munk, W. H. & Wunsch, C. *Deep-Sea Res.* **45**, 1977–2010 (1998).
3. Gregg, M. C. J. *Geophys. Res.* **94**, 9686–9698 (1989).
4. Ledwell, J. R., Watson, A. J. & Law, C. S. *J. Geophys. Res.* **103**, 21499–21529 (1998).
5. Polzin, K. *et al. Science* **276**, 93–96 (1997).
6. Wunsch, C. *Nature* **405**, 743–744 (2000).
7. Toggweiler, J. R. & Samuels, B. J. *Phys. Oceanogr.* **28**, 1832–1852 (1998).
8. Döös, K. & Coward, A. C. *Int. WOCE Newsl.* **27**, 3–4 (1997).
9. Foster, T. D. & Carmack, E. C. *Deep-Sea Res.* **23**, 301–317 (1976).
10. Schmitz, W. J. *Rev. Geophys.* **33**, 151–173 (1995).

Antibiotic resistance

How wild are wild mammals?

In bacteria associated with humans, antimicrobial resistance is common, both in clinical isolates and in the less-studied commensal flora, and it is thought that commensal and environmental bacteria might be a hidden reservoir of resistance. Gilliver *et al.* have reported that resistance is also prevalent in faecal bacteria from wild rodents living in northwest England¹. Here we test the faeces of moose, deer and vole in Finland and find an almost complete absence of resistance in enterobacteria. Resistance is thus not a universal property of enterobacterial populations, but may be a result of the human use of antibiotics.

Bacterial resistance to antimicrobial agents has become a serious problem in modern medicine and a debated evolutionary question². The use — and misuse — of antibiotics is generally blamed, but it has also been claimed that there must be other reasons for the increase in resistance³. This question is important: if resistance increases independently of antibiotic use, restrictive policies would be unnecessary. One way to test the effect of human activities is to compare the resistance frequencies of popula-

Table 1 Antimicrobial resistance in enterobacteria from the faeces of wild moose, deer and vole

Bacteria (no. of isolates)	Percentage resistant (MIC ₅₀ , MIC ₉₀)						
	AMP ≥32*	AMC ≥32/16	CEX ≥32	CXM ≥32	CTX ≥64	ATM ≥32	IPM ≥16
<i>Escherichia coli</i> (98)	0 (4; 8)	0 (4; 4)	0 (8; 8)	0 (4; 4)	0 (0.06; 0.06)	0 (≤0.06; 0.1)	0 (≤0.2; ≤0.2)
<i>Enterobacter agglomerans</i> group (48)	NA (8; 16)	NA (4; 4)	NA (8; 16)	4 (4; 4)	0 (0.06; 0.1)	0 (≤0.06; 0.1)	0 (0.5; 1)
<i>Yersinia</i> spp. (29)	NA (32; 64)	NA (32; >64)	NA (>64; >64)	7 (4; 4)	0 (0.2; 0.5)	0 (0.5; 0.5)	0 (0.5; 1)
<i>Serratia</i> spp. (11)	NA (32; 128)	NA (8; >64)	NA (>64; >64)	82 (64; >64)	0 (0.5; 1)	0 (0.5; 0.5)	0 (0.5; 0.5)
Bacteria (no. of isolates)	Percentage resistant (MIC ₅₀ , MIC ₉₀)						
	GEN ≥16	STR ≥32	NAL ≥32	CIP ≥4	CHL ≥32	TET ≥16	TMP ≥16
<i>Escherichia coli</i> (98)	0 (0.5; 0.5)	1 (4; 4)	0 (2; 4)	0 (0.03; 0.06)	0 (4; 8)	0 (1; 2)	0 (0.2; 0.5)
<i>Enterobacter agglomerans</i> group (48)	0 (0.2; 0.2)	0 (2; 2)	0 (1; 4)	0 (0.03; 0.1)	0 (≤2; ≤2)	0 (1; 1)	0 (≤0.06; 0.1)
<i>Yersinia</i> spp. (29)	0 (0.2; 0.5)	0 (2; 4)	0 (1; 1)	0 (0.01; 0.03)	0 (8; 8)	0 (2; 2)	0 (1; 2)
<i>Serratia</i> spp. (11)	0 (0.2; 0.5)	0 (2; 4)	0 (2; 2)	0 (0.06; 0.1)	0 (4; 16)	NA (2; 64)	0 (0.2; 0.5)

MIC₅₀ and MIC₉₀ are the antibiotic concentrations (mg l⁻¹) at which 50 and 90%, respectively, of the tested population is inhibited from growing. AMP, ampicillin; AMC, amoxycillin/clavulanic acid; CEX, cephalothin; CXM, cefuroxime; CTX, cefotaxime; ATM, aztreonam; IPM, imipenem; GEN, gentamicin; STR, streptomycin; NAL, nalidixic acid; CIP, ciprofloxacin; CHL, chloramphenicol; TET, tetracycline; TMP, trimethoprim; SUL, sulphamethoxazole. NA, not applicable: most strains or species intrinsically resistant.

*Resistance breakpoint (mg l⁻¹) according to the National Committee for Clinical Laboratory Standards.

tions of the same bacterial species that have or have not been exposed to humans.

Fresh faeces from newly felled moose (*Alces alces*; *n* = 16) and white-tailed deer (*Odocoileus virginianus*; *n* = 7) were collected in the autumn of 1999 by hunters in two areas of Uusimaa, southern Finland. Faecal pellets were prepared from bank voles (*Clethrionomys glareolus*; *n* = 23) trapped in Ostrobothnia, western Finland⁴; these had been stored whole at -20° C for less than one year. Five bacterial colonies per sample, representing all different colonial morphologies present, were identified to at least genus level, and the minimum inhibitory concentrations (MICs) of 15 antibiotics were determined, as previously described⁵.

The ungulate faecal flora was similar to human flora, with *Escherichia coli* as the main species. In vole faeces, *Enterobacter agglomerans* and *Yersinia* spp. dominated. Results are given only for genera represented by more than four isolates. The only resistance found was to cefuroxime (Table 1) and to streptomycin (in one sample of *E. coli*; this could be transferred by conjugation⁶ to *E. coli* C600). Most of the cefuroxime resistance was, as judged from MIC profiles, most likely caused by a class A (Bush group 2e) cefuroximase similar to the chromosomal *Proteus vulgaris* enzyme⁷, and was thus most probably indigenous. It was found not to be caused by the most common transferable class-A β-lactamases: the cerufoxime-resistant strains were tested by using the polymerase chain reaction for the presence of TEM and SHV⁸, but only one strain contained a TEM-type enzyme and none carried SHV.

These results disagree with those from a

study of enterobacteria from wild English rodents, where extremely high resistance was found¹. The English study questions the usefulness of restricting antibiotic use, as these rodents are presumed to have had no contact with antibiotics. However, the overall load from antibiotic use in England is larger than in Finland: the mean number of inhabitants per square kilometre in Finland is 17, compared to 378 in England (see www.statistics.gov.uk and www.stat.fi); also, the load from agriculture is less in Finland — there are, for example, ten times fewer cattle and five times fewer pigs than in the UK (see www.stat.fi and www.maff.gov.uk). Since 1996, the use of antibiotic additives in animal feed has gradually been abandoned, but occasional contact cannot be ruled out. Our sampled populations almost certainly represent wild animal populations better.

In faecal flora, *E. coli* is the species showing the most resistance (resistance to streptomycin, to sulphamethoxazole, and to tetracycline is highest at 14–18%, even in healthy people)⁵. Resistance is known to increase with increased exposure to antibiotics and during hospitalization⁵. Our finding of an *E. coli* population that has never been exposed to humans and which is free of resistance to antibiotics strongly suggests that the widespread resistance found in all *E. coli* populations associated with humans must be caused by human activities. Antibiotic restrictions whenever feasible are still very much on the agenda.

Monica Österblad*, Kai Norrdahl†, Erkki Korpimäki†, Pentti Huovinen*

*Antimicrobial Research Laboratory, National Public Health Institute, Kiinamyllynkatu 13, FIN-20520 Turku, Finland

†Section of Ecology, Department of Biology, University of Turku, FIN-20014 Turku, Finland

- Gilliver, M., Bennett, M., Begon, M., Hazel, S. & Hart, C. *Nature* **401**, 233–234 (1999).
- Baquero, F. & Blázquez, J. *Trends Ecol. Evol.* **12**, 482–487 (1997).
- Levy, S. B., Marshall, B., Schluederberg, S., Rowse, D. & Davis, J. *Antimicrob. Agents Chemother.* **32**, 1801–1806 (1988).
- Korpimäki, E. & Norrdahl, K. *Ecology* **79**, 2448–2455 (1998).
- Österblad, M. *et al.* *Antimicrob. Agents Chemother.* **44**, 1479–1484 (2000).
- Murray, B. E. & Hodel-Christian, S. L. in *Antibiotics in Laboratory Medicine* (ed. Lorian, V.) 453–501 (Williams & Wilkins, Baltimore, 1996).
- Livermore, D. M. & Williams, J. D. in *Antibiotics in Laboratory Medicine* (ed. Lorian, V.) 502–578 (Williams & Wilkins, Baltimore, 1996).
- Manninen, R. *et al.* *Clin. Microbiol. Infect.* **3**, 408–413 (1997).

Gilliver et al. reply — The study by Österblad *et al.* confirms the importance of understanding the role of commensal bacteria, particularly in wildlife, in the ecology of antibiotic resistance. The two studies combined suggest that the gut flora of wildlife populations with very little or no contact with either humans or anthropogenic antibiotics (Österblad *et al.*'s study) may have negligible levels of antibiotic resistance, whereas wildlife populations living in closer proximity to humans but still with no known direct contact with anthropogenic antibiotics (our study) may have much higher levels of antibiotic resistance.

These conclusions fit with earlier findings of a higher prevalence of antibiotic resistance among baboons living close to humans than in baboons in more isolated populations¹. Questions that still need to be addressed concern the extent and frequency of antibiotic exposure necessary to generate significant resistance, what determines the dynamics of decline in resistance following restrictions in antibiotic use, and the nature and extent of any reservoir of antibiotic resistance that may exist in natural environments and which could undermine future attempts to manage resistance.

These questions can only be resolved by thorough spatial and temporal mapping of antibiotic resistance in natural environments. We inferred from our study that it would be unwise to assume that resistance would decline significantly as a consequence of restricted use of antibiotics. This suggestion still holds, because resistance has been maintained for over three years at our study site, over several generations of rodents, without any obvious exposure to antibiotics.

We agree with Österblad *et al.* that antibiotic restrictions should still be very much on the agenda, but that agenda must include concerted attempts to understand what the consequences of restrictions are likely to be.

Maira A. Gilliver, Malcom Bennett, Michael Begon, Sarah M. Hazel, C. Anthony Hart

Centre for Comparative Infectious Diseases, University of Liverpool, PO Box 147, Liverpool L69 3BX, UK

- Routman, E., Miller, R. D., Phillips-Conroy, J. & Hartl, D. L. *Appl. Environ. Microbiol.* **50**, 749–754 (1985).

X-ray clusters of galaxies as tracers of structure in the Universe

Stefano Borgani*† & Luigi Guzzo‡

* INFN, Sezione di Trieste, c/o Dipartimento di Astronomia dell'Università, Via Tiepolo 11, I-34131, Trieste, Italy

† INFN, Sezione di Perugia, c/o Dipartimento di Fisica dell'Università

‡ Osservatorio Astronomico di Brera, Via Bianchi 46, I-23807, Merate, Italy

Clusters of galaxies are visible tracers of the network of matter in the Universe, marking the high-density regions where filaments of dark matter join together. When observed at X-ray wavelengths these clusters shine like cosmic lighthouses, as a consequence of the hot gas trapped within their gravitational potential wells. The X-ray emission is linked directly to the total mass of a cluster, and so can be used to investigate the mass distribution for a sizeable fraction of the Universe. The picture that has emerged from recent studies is remarkably consistent with the predictions for a low-density Universe dominated by cold dark matter.

The present-day appearance of the galaxy distribution is the result of the gravitational growth of the initial fluctuations laid down during the very early stages of cosmic history, coupled to dissipative processes which lit up matter and made it visible to our telescopes. Cosmological models directly predict the distribution and gravitational growth of the mass, not of the light we actually observe from galaxies. Thus, the challenge of unveiling the nature of the cosmic initial conditions from the observed structure of the Universe is a process which is ideally characterized by two steps. The first involves constructing maps of the distribution of luminous objects on sufficiently large scales to encompass a representative portion of the whole Universe. The second requires relating such 'light maps' to the underlying mass through a physically motivated and robust recipe, the two distributions being, in principle, not related by a one-to-one correspondence. Although galaxy surveys are now able to probe the distribution of luminous structures over scales larger than 100 megaparsecs (Mpc, 1 pc = 3.26 light years), their relation to the actual mass distribution depends on the still poorly understood physics of galaxy formation and evolution.

Clusters of galaxies, the most evident concentrations in galaxy maps, can themselves be used as tracers of the large-scale structure of the Universe. Although they miss the fine details, they can be used efficiently to study extremely large volumes at a reduced cost in terms of telescope time with respect to fully sampled galaxy surveys. Already the earliest statistical samples of clusters from visually compiled catalogues^{1–3} reached typical depths of a few hundreds of Mpc, and the Abell catalogue in particular¹ still represents one of the main resources for cosmological studies^{4,5}. It is within the "gravitational sinks" of galaxy clusters that evidence for the elusive dark matter was first found in the 1930s (refs 6, 7), as a necessary ingredient to explain the fast motions observed for cluster galaxies.

The loose definition of a cluster as a collection of galaxies is however intrinsically uncertain, and is certainly not optimal for estimating its mass, as is required for linking observations and theoretical predictions. Fortunately, about 20–30% of the optically invisible mass of a cluster is in the form of a diffuse hot gas⁸, trapped and heated to a temperature of the order of 10^8 K by its gravitational potential. At such high temperatures, this gas is a fully ionized plasma, producing powerful X-ray emission by free–free electron–ion interactions, the so-called bremsstrahlung radiation. With total luminosities of $\sim 10^{43}$ – 10^{45} erg s^{−1}, and large physical dimensions (~ 1 Mpc), galaxy clusters can be recognized over the rather sparse X-ray sky⁹ as extended sources, out to very large distances. The X-ray luminosity is shown to correlate well with the cluster mass; indeed, it provides a fairly direct way to make a robust comparison

of the observed clustering with the predictions of cosmological models. The most recent projects to study large-scale structure and benefiting from these advantages^{10–14} are based on the all-sky survey by the ROSAT satellite and have thus far accumulated distances for $\sim 1,000$ clusters within a volume of the order of 10^8 – 10^9 Mpc³. The degree of clumpiness observed in these samples on scales of several hundreds of Mpc is remarkable, and is larger than previously indicated by galaxy surveys. The amplitude and shape of the corresponding distribution of mass inhomogeneities, which for X-ray clusters can be precisely computed, suggest a Universe dominated by non-baryonic cold dark matter, whose density is about one-third of that needed for it to recollapse under its own self-gravity. Coupled to the recent strong evidence for a nearly flat spatial geometry from the anisotropy of the cosmic microwave background (CMB)^{15,16}, this reinforces the apparent need for an extra cosmic energy, that is, what is commonly associated with the cosmological constant.

Large-scale view of the Universe

Modern cosmology uses the distribution of galaxies and clusters to map the structure of the Universe on scales which have presently reached a few hundred h^{-1} Mpc (here h is the Hubble constant, H_0 , in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$; see Box 1). Systematic redshift surveys of galaxies first started in the mid-1970s (ref. 17) and became widespread during the past two decades (see ref. 18 and references therein; also, see Box 1 for the definition of redshift). The main features of the large-scale distribution of galaxies emerging from these cosmic maps include long, thin superclusters, with stronger concentrations—rich clusters—located at their intersections, surrounding large regions mainly devoid of galaxies. This is shown in Fig. 1a, where we plotted the distribution of about 26,000 galaxies composing the Las Campanas redshift survey (LCRS, blue points)¹⁹. The evident inhomogeneity in the distribution of galaxies can be statistically quantified at the simplest level in terms of their two-point correlation function $\xi(r)$, which measures the excess probability, with respect to a random distribution, of observing a pair of galaxies separated by a distance r (ref. 20) (see Box 1). The squares in Fig. 2 show the correlation function measured from the LCRS galaxy survey²¹. Consistent with other samples of normal galaxies, $\xi(r)$ is well described by a power law $(r/r_0)^{-\gamma}$ with $r_0 \approx 5 h^{-1}$ Mpc and $\gamma \approx 1.8$ for separations r between about 0.1 and $10 h^{-1}$ Mpc. This tells us, roughly, that galaxies are strongly clustered within this scale range, while tending to a more homogeneous distribution at larger separations^{22–24}.

This variety of observed cosmic structures is interpreted as arising from the gravitational growth of initially tiny fluctuations,

generated in the early stages of the life of the Universe^{25,26}. Cosmological models yield predictions for the character of such initial fluctuations in terms of their power spectrum $P(k)$, that is, the Fourier transform of $\xi(r)$ (see Fig. 3 and Box 1). However, their connection to the observations is not trivial. Although direct and reliable predictions can be made about the distribution of matter concentrations of a given mass²⁷, it is far less straightforward to

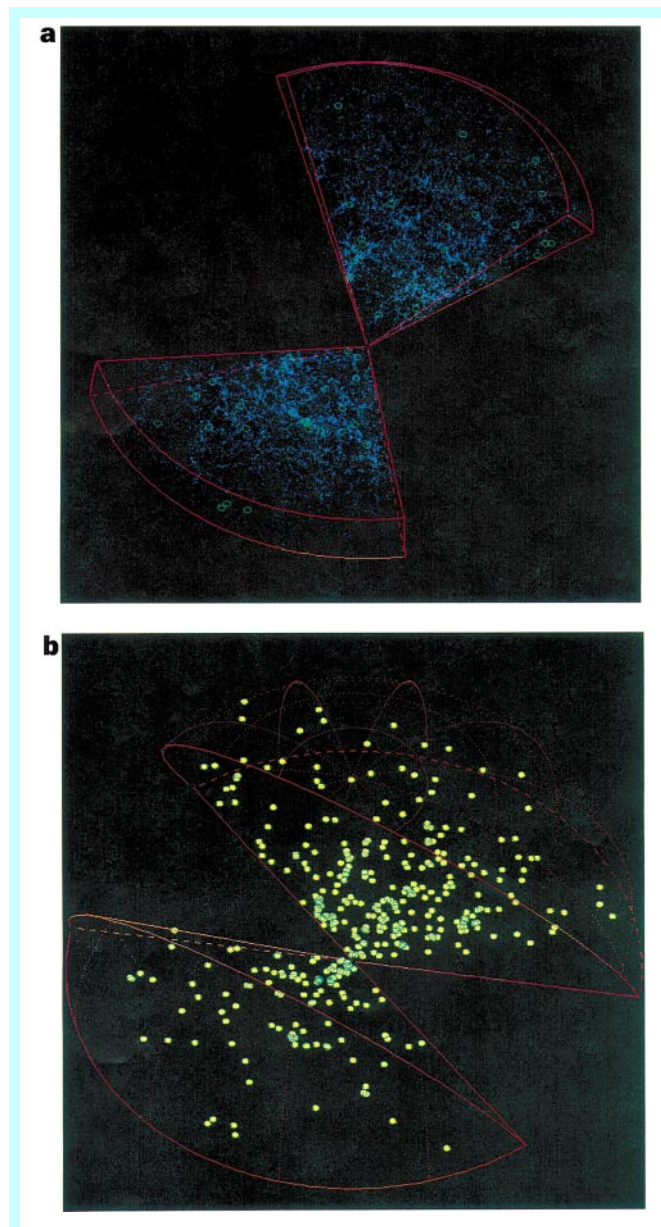


Figure 1 Maps of the cosmic web. **a**, The distribution of ~26,000 galaxies in the northern and southern slices of the Las Campanas redshift survey¹⁹ (LCRS), to a maximum distance of 600 h^{-1} Mpc from the observer, located in the centre. The typical morphology of large-scale structure, with filaments and voids, is evident. Superimposed on the 'blue dust' of the galaxy distribution, the green circles mark the positions of those X-ray clusters of galaxies from the REFLEX survey¹³ that lie approximately within the volume of the LCRS. **b**, The full volume of the REFLEX survey (which contains all clusters brighter than an X-ray flux of 3×10^{-12} erg s $^{-1}$ cm $^{-2}$ over a large part of the southern sky) within the same distance limit (similar orientation to **a**, south is up for display purposes). The much larger volume sampled by clusters and their clustering along filamentary structures is evident. The missing part of the hemisphere corresponds to the region highly obscured by the Milky Way disk ($\pm 20^\circ$ in galactic latitude). Similar volumes will be partly filled by galaxy redshift measurements only in the coming years^{77,78}.

Box 1

A large-scale structure primer

The redshift z of a galaxy is defined as the fractional increase in the observed wavelength of the emitted radiation with respect to its laboratory value: $z = (\lambda - \lambda_0)/\lambda_0$. In the standard cosmological model this is interpreted to be due to the global expansion of the Universe, and to first approximation (that is, neglecting any acceleration of the cosmic expansion) it can be thought of as the consequence of a recession velocity $v_{\text{rec}} = cz$ (where c is the speed of light), proportional to the distance to the galaxy itself, $v_{\text{rec}} \approx H_0 d$. This is the empirically verified Hubble law and the constant of proportionality H_0 is the Hubble constant. Through this relation we can measure the distance to faraway galaxies simply by observing the redshift in their spectrum.

According to the theory of gravitational instability, cosmic structures arise from gravitational growth on initially tiny inhomogeneities, generated in the very early stages of the Universe. These cosmic inhomogeneities are described by the fluctuation field $\delta(\mathbf{x}, z) = (\delta\rho/\rho)_{\mathbf{x},z}$, where $\rho(\mathbf{x}, z)$ is the density of the Universe at the point $\mathbf{x}$ and at redshift z . A basic statistical characterization of $\delta(\mathbf{x})$ is represented by its two-point correlation function, $\xi(r)$, which describes the excess fluctuation with respect to a uniform distribution between two points at separation r . The Fourier transform of $\xi(r)$ is the power spectrum:

$$P(k) = \frac{1}{2\pi^2} \int_0^\infty dr \xi(r) \frac{\sin kr}{kr} \quad (1)$$

The power spectra of two representative cosmological models are compared in Fig. 2 to observational results. Even keeping fixed the assumption of a Universe dominated by Cold Dark Matter (CDM), different shapes for the matter $P(k)$ can be obtained by varying the values of Ω_m , h and the amount of baryons⁷². Its amplitude is instead usually determined in a phenomenological way, by matching some observed measure of the root mean squared (r.m.s.) fluctuation (in the mass, not in the light) at some scale. One way is to reproduce the level of CMB anisotropy measured by the COBE satellite⁶⁴, that probes fluctuations on scales of $\sim 10^3$ Mpc. On a completely different range of scales, an alternative method is offered by the local abundance of galaxy clusters, by virtue of their ability to probe the mass fluctuations. Because they formed from the collapse of density fluctuations on a scale of $\sim 10 h^{-1}$ Mpc, their number density today is directly proportional to the r.m.s. amplitude of such fluctuations at the onset of their growth. As massive clusters arise from rare high peaks of the density fluctuation field, their number density is sensitive to the amplitude of these fluctuations, providing a precise constraint on the amplitude of $P(k)$ ^{65,73–75}. For instance, the simulations shown in Fig. 6 refer to two models having different Ω_m but normalization-tuned so as to produce a comparable number of clusters at $z = 0$.

Once $P(k)$ is normalized at the present time, its amplitude in the past depends on the evolution of density fluctuations. As long as fluctuations are small, their growth is independent of the position, so that $\delta(\mathbf{x}, z) = D(z)\delta(\mathbf{x}, z = 0)$. The redshift dependence of $D(z)$ is determined by comparing the timescales for gravitational collapse of a perturbation and for the expansion of the Universe^{25,26}. These two timescales are always identical when $\Omega_m = 1$, thus implying that the perturbation growth equals the cosmic expansion factor: $D(z) = (1 + z)^{-1}$. For a low-density Universe, the two timescales are similar at early epochs, whereas the cosmic expansion overtakes the fluctuation growth when expansion is no longer influenced by the self-gravity of the Universe. This happens at an epoch corresponding to $1 + z \approx \Omega_m^{-1/3}$ or $1 + z \approx \Omega_m^{-1}$ in the cases with and without a cosmological constant providing flat geometry, respectively. Therefore, in a low- Ω_m Universe the fluctuation growth progressively slows down at later epochs and eventually freezes.

The plots in Fig. 6 show the different degrees of evolution characterizing models with different values of Ω_m , as indicated by the different number of clusters already formed at high redshift. For this reason, the determination of the number density of distant clusters is a powerful indicator for cosmological models (see refs 51, 76 and references therein).

describe the clustering properties of the galaxies which have then formed within these potential wells. To relate their observable characteristics (such as luminosity, colour and morphology) to their actual total mass, would require a complete understanding of all physical processes regulating the formation and evolution of stars within galaxies. Although much work is currently dedicated to such understanding (see refs 28, 29 and references therein), this is far from complete.

Clusters as tracers of the cosmic web

Once we accept the idea of luminous objects as tracers of the underlying mass structure of the Universe, we might think of alternatives to using single galaxies. Figure 1b plots the distribution of a large sample of clusters of galaxies^{13,30} detected by the X-ray satellite ROSAT, one typical example of which is shown in the combined optical–X-ray picture of Fig. 4. Comparison of Fig. 1a and b shows how although massive clusters provide a coarser mapping of large-scale structure, they are very efficient for sampling extremely large ($r > 100 h^{-1}$ Mpc) volumes; this translates into a more modest investment of telescope time to reconstruct their three-dimensional distribution. In general, rich clusters such as shown in Fig. 4 contain several hundreds of galaxies within a typical size of $\sim 1 h^{-1}$ Mpc, and represent the largest gravitationally bound structures in the Universe. Galaxies in clusters are observed to move along their orbits with a typical line-of-sight velocity dispersion $\sigma_v \approx 1,000 \text{ km s}^{-1}$, so that the time required for a galaxy to cross the cluster is approximately $t_{\text{cr}} = r/\sigma_v \approx 10^9 \text{ yr}$. This means that, given the typical age of the Universe, $t_U \approx H_0^{-1} \approx 10^{10} h^{-1} \text{ yr}$, rich clusters have had enough time to evolve into dynamically relaxed systems. While a continuous infall of galaxies moving along radial orbits takes place in the cluster outskirts, the central regions have reached virial equilibrium, in which the total mass is related to the galaxy velocity dispersion as $M_{\text{vir}} \approx 3\sigma_v^2 r/G \approx 10^{15} h^{-1} M_\odot$, where $M_\odot$ is the mass of the Sun. Cluster masses measured in this way^{6,7} are about ten times larger than was expected from the sum of the masses of the individual member galaxies. This observation represented the

earliest evidence for a ‘hidden’ form of matter permeating the Universe.

In Figs 2 and 3 we plotted the correlation function and power spectrum measured for clusters selected through their X-ray emission^{31,32}. Both figures show how clusters display a clustering amplitude significantly larger than galaxies, while maintaining a similar shape: the correlation length of X-ray luminous clusters is roughly four times larger: $r_0 \approx 20 h^{-1} \text{ Mpc}$. This shows explicitly how changing the kind of tracer used allows us to measure different clustering properties. When first observed in optically selected samples^{33–35}, this motivated the concept of bias in the distribution of cosmic structures³⁶; the strong clumpiness of the cluster distribution is just the natural consequence of clusters tracing only the high-density peaks of the underlying mass density field. In general, the word ‘bias’ is meant to indicate the relation between the distribution of a given class of observable objects, such as galaxies and clusters, and the underlying distribution of matter: what regulates the amount of clustering of a class of objects is just the characteristic mass of the objects themselves. In the case of clusters, their correlation function and power spectrum are predicted to be amplified with respect to those of the matter distribution, by a constant biasing factor, whose value is uniquely known once the cluster mass and the cosmological model are specified²⁷. This has the important consequence that, given a sample of clusters selected according to their mass, the observed clustering is a direct test of the cosmological model. Although clusters are clearly more biased tracers of the mass distribution with respect to galaxies, they are more useful because their biasing factor can easily be computed.

This simple theory is complicated by the loose visual definition of a cluster as an agglomerate of galaxies. The traditional way of selecting clusters is in fact based on the ‘eyeball’ detection of overdensities in the projected galaxy distribution on the sky¹. To estimate the cluster mass, richness is measured by counting, within a fiducial aperture radius, the number of galaxies which belong to the cluster. However, partly because of the difficulty of properly correcting for the contamination by background and foreground

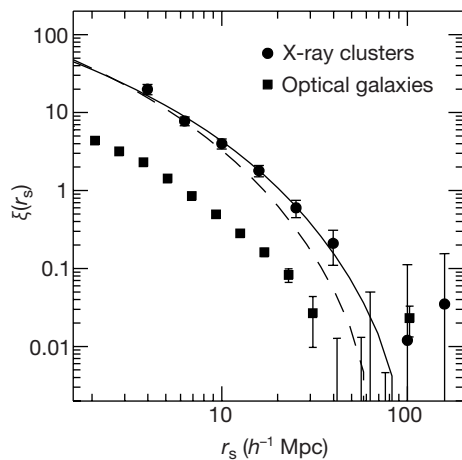


Figure 2 Statistical description of clustering. The two-point correlation functions ξ of galaxies (squares) and X-ray clusters of galaxies (circles), computed from the two data sets of Fig. 1^{21,31}, plotted as a function of separation r_s (where the subscript s indicates that all object distances are computed from the measured redshift). The curves are the predictions of two CDM models, with different density parameters Ω_m , for an X-ray cluster survey with the same flux limit as the real data. Solid curve: $\Omega_m = 0.3$ and Hubble parameter $h = 0.7$; dashed curve: $\Omega_m = 0.5$ and $h = 0.6$. Both models have flat spatial geometry provided by a cosmological constant contribution (that is, $\Omega_m + \Omega_\Lambda = 1$) and power-spectrum normalization chosen so as to be consistent with measured CMB anisotropies^{15,16,64}. Also, we take $\Omega_{\text{bar}} h^2 = 0.019$ for the baryon density⁷⁹.

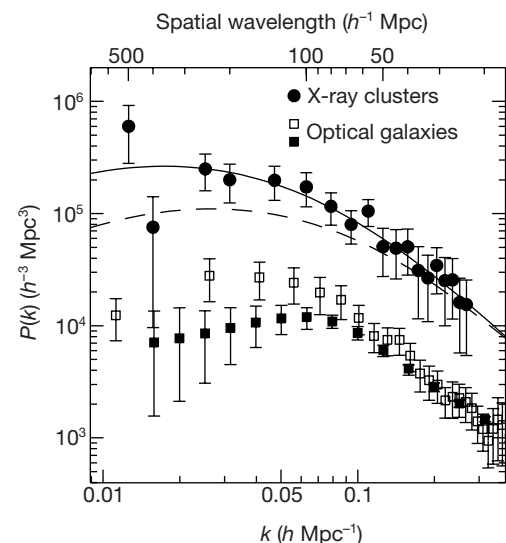


Figure 3 The power spectrum of the distribution of galaxies and X-ray clusters of galaxies from the data of Fig. 1. The squares are from galaxy data, which in addition to the LCRS points (filled squares⁸⁰) include a measure from another survey with better volume coverage (open squares⁸¹). The filled circles show an estimate of the power spectrum of X-ray selected clusters from the REFLEX survey³². We note the rather different amplitude between the galaxy and cluster power spectra, similarly to that shown by correlation functions (see Fig. 2). The two curves are theoretical predictions from the same cosmological models shown in Fig. 2.

galaxy counts^{37,38}, the richness is itself an intrinsically poor mass indicator compared to X-ray luminosity (see Fig. 5). This highlights the difficulty of using the distribution of clusters selected by optical richness to quantitatively constrain theoretical scenarios for the formation of cosmic structures.

Clusters in X-rays

The first X-ray observations of nearby galaxy clusters^{39–41} showed that they are in general associated with extended X-ray sources⁴² with luminosities L_X in the range $\sim 10^{43}–10^{45}$ erg s^{−1}, whose emission originates in a diffuse hot intergalactic medium permeating the cluster potential well (see ref. 43 for a historical review). At equilibrium gas and galaxies in the cluster have to share the same dynamics, so we expect to measure a gas temperature $k_B T \approx \mu m_p \sigma_v^2$, where m_p is the proton mass, $\mu \approx 0.6$ is the gas mean molecular weight and k_B is Boltzmann's constant. Given the typical cluster velocity dispersions, this corresponds to a temperature of a few keV, as is indeed measured by X-ray observations. At such energies the intracluster medium (ICM), composed mainly of hydrogen, behaves like a fully ionized plasma with an atomic density of $\sim 10^{-3}$ particles per cm³. The scattering between free electrons and ions in such conditions produces thermal bremsstrahlung radiation, which peaks in the X-ray region of the electromagnetic spectrum. For this mechanism, the emissivity (that is, the energy released per unit time, frequency and volume) at frequency ν and temperature T is: $\epsilon_\nu \propto n_e n_i T^{-1/2} \exp(-h_p \nu / k_B T)$, where n_e and n_i are the number densities of electrons and ions, respectively, and h_p is the Planck constant. This expression shows why the identification of

clusters in the X-ray band is less affected by projection effects with respect to the optical selection based on galaxy overdensities: whereas the optical emission depends linearly on the number of galaxies, the X-ray emission depends on the square of the local gas density, making clusters stand out more sharply in the X-ray than in the optical light. The contours in Fig. 4 explicitly show that in the X-ray band the cluster emerges as a single extended source.

The X-ray luminosity of galaxy clusters is a direct function of the total cluster mass, at least in a phenomenological way⁴⁴, as demonstrated by the plot in the bottom panel of Fig. 5. This is further supported by the observation of a well defined relation between the X-ray luminosity and the temperature of the ICM, the latter being a reliable indicator of the depth of the cluster gravitational potential^{45,46}. Such a relationship has been observationally calibrated at low redshift with a fairly large number of clusters, showing that $L_X \propto T^\alpha$ with $\alpha \approx 3$, a small scatter of $\lesssim 30\%$ around this relation^{47,48}, and no evidence for evolution out to $z \approx 1.3$ ^{49–51,83}. These observations have profound implications for the physics of the ICM, as the steep slope of the luminosity–temperature relation and the lack of evolution cannot be accounted for by the action of gravity only. Additional physical mechanisms, such as radiative cooling in the central cluster regions and heating from supernovae explosions and active galactic nuclei, seem to influence the thermodynamics of the intracluster gas (see refs 52, 53 and references therein). We note that, whichever mechanism determines the ICM



Figure 4 The visual and X-ray appearance of the rich galaxy cluster RXCJ1206.2-0848 at redshift $z = 0.44$, discovered by the REFLEX survey³⁰. The optical image has been obtained by combining three images taken with the European Southern Observatory (ESO) 3.6-m telescope, corresponding to the red, green and blue bands. The yellowish colour of objects in the central region reflects the old stellar population characterizing elliptical galaxies, that typically dominate in cluster cores. We note the presence of at least two probable gravitational arcs (with blue colour), near the central giant elliptical galaxy. These are the result of the gravitational lensing phenomenon, by which the cluster mass distorts and amplifies the images of background galaxies (thus providing an independent way to probe its potential⁸²). The contours show the X-ray emission from the cluster measured by the ROSAT all-sky survey, which yield a luminosity $L_X \approx 5 \times 10^{44} \text{ h}^{-2} \text{ erg s}^{-1}$. The field of view in this image is about 5 arcmin, corresponding to a physical size of $\sim 1 \text{ h}^{-1} \text{ Mpc}$ at the cluster distance.

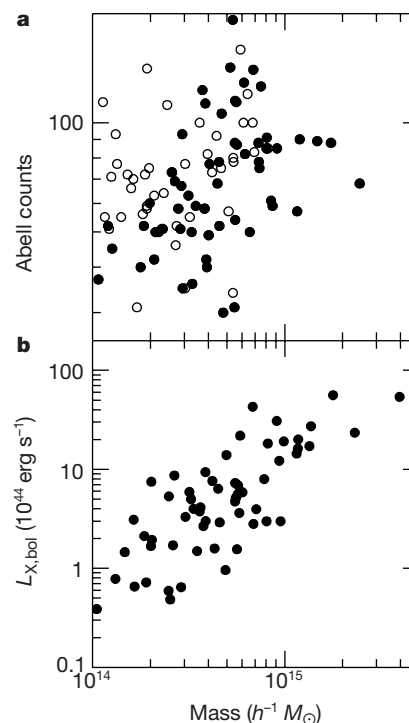


Figure 5 Correlations between mass and observational properties of galaxy clusters. The correlations with Abell^{1,2} richness counts N_A (**a**), and their X-ray luminosity $L_{X, \text{bol}}$ (**b**) are based on the merge of a compilation of clusters with accurate measures of velocity dispersion⁸³ and a sample of X-ray bright Abell clusters (XBACs; ref. 10). Cluster masses are estimated by applying the virial theorem to the velocity dispersions of member galaxies. In **a** and **b**, filled circles represent those clusters belonging to XBACs, and open circles those Abell clusters that have too low an X-ray emission to be included in XBACs. The weak correlation in **a** demonstrates the difficulty of using clusters selected by optical richness to strongly constrain cosmological scenarios. This aim is better achieved with clusters selected by X-ray luminosity, which clearly shows a tighter correlation with the cluster mass.

thermodynamics, the very fact that gas temperature and luminosity are observed to be closely correlated demonstrates that L_X is indeed a reliable indicator of cluster mass.

Besides X-ray imaging, radio observations of the Sunyaev–Zel'dovich (SZ) effect are becoming an important alternative method of detecting distant clusters with a well defined mass selection (see refs 54 and 84 for recent reviews). The SZ effect is observed as a surface brightness variation in the CMB in the direction of a galaxy cluster and is produced by the scattering of CMB photons on the energetic electrons of the ICM (known as inverse Compton). Owing to its nature, the SZ signal depends only on the total thermal energy of the ICM and is less sensitive than X-ray emission to its detailed structure and complexity. The SZ signal has been now detected for several known clusters, but no systematic blind search based on this effect has yet been realized. The extensive use of this technique for large-scale structure studies will become feasible only with the next generation of high-resolution microwave surveyors, such as the Planck satellite.

Clustering of X-ray clusters and implications for cosmology

A quantitative measure of the clustering of X-ray clusters has been possible only in recent years⁵⁵. An important advance has been the ROSAT all-sky survey (RASS; ref. 56), which for the first time provided a X-ray imaging survey of the whole sky, in which thousands of clusters as faint as $f_X = 10^{-12} \text{ erg s}^{-1} \text{ cm}^{-2}$ can be detected. (Depending on the technology of their mirrors and

detectors, different X-ray satellites cover different energy ranges, which need to be specified when quoting fluxes and luminosities. The ROSAT band used here covers the energy range 0.1–2.4 keV.)

Early studies of X-ray clusters identified within subregions of the RASS produced first estimates of the clustering of these objects^{57–60,86}. The quality of such X-ray cluster catalogues has substantially improved through the recent completion of the optical identification and redshift measurement for a statistically complete sample of nearly 500 RASS clusters over half of the sky^{13,30} (see also Fig. 1). Because of the precision intrinsic to the X-ray selection, these data are now providing an unprecedented large-scale description of the structure of the Universe and a powerful testbed for cosmological scenarios. Figures 2 and 3 highlight the accuracy with which the clustering of clusters is now measured on intermediate scales ($\sim 10\text{--}10 h^{-1} \text{ Mpc}$) by the correlation function³¹ and on the largest scales achievable to date ($\sim 500 h^{-1} \text{ Mpc}$) with the power spectrum³². Extension of this work to the whole sky¹⁴ and to fainter fluxes will eventually result in a complete sample of about 1,500 clusters at a median redshift of $z \approx 0.1$.

Given the well defined X-ray selection function for these samples, the interpretation of the observed clustering in terms of cosmological models is now relatively straightforward. Once the gravitational growth of cluster-sized mass clumps within a given model is predicted, either analytically^{27,61} or numerically^{62,63}, then the corresponding X-ray emission from the ICM can be computed using the mass–temperature–luminosity relation discussed above. Figure 6 is

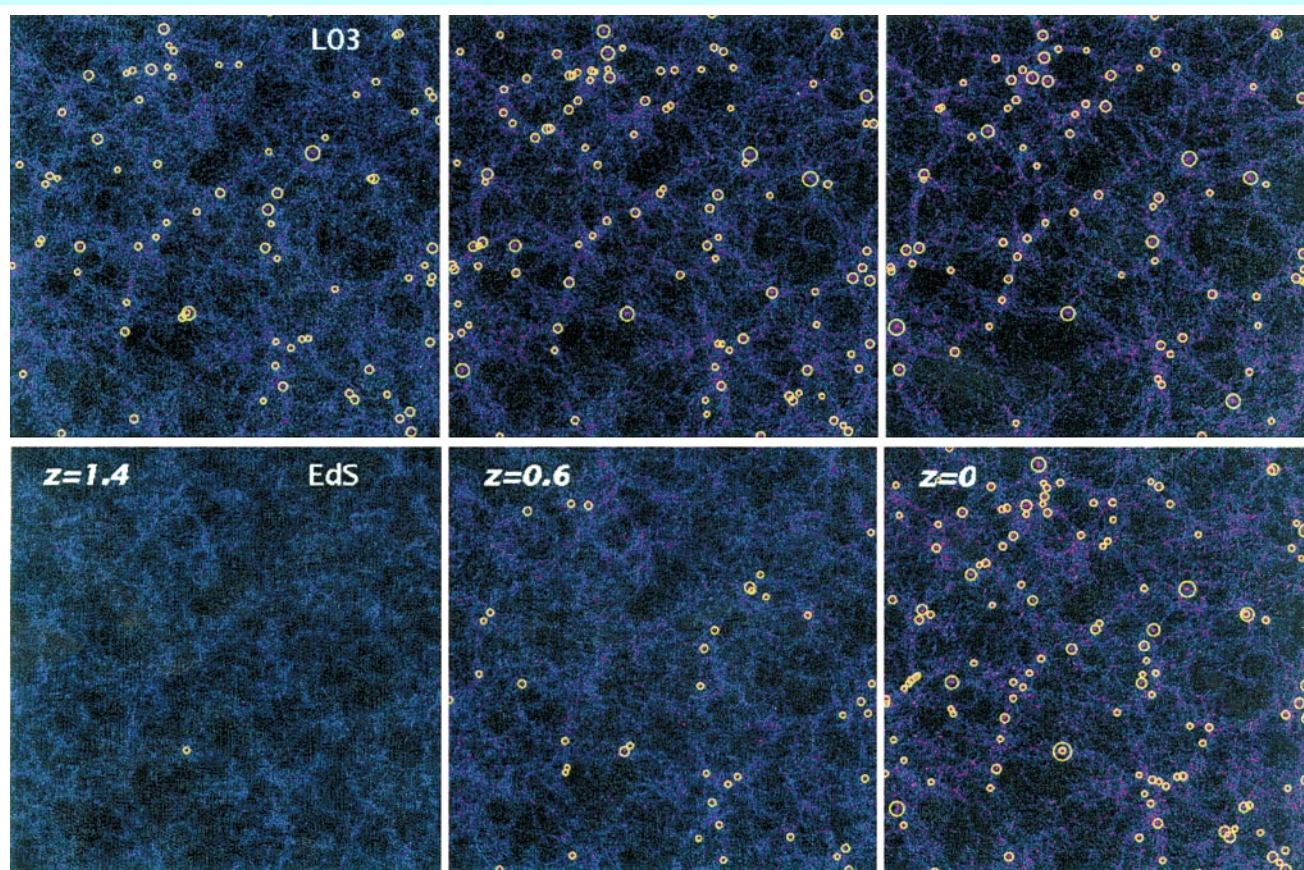


Figure 6 The evolution of gravitational clustering simulated using an N-body code for two different models. Each of the three redshift snapshots shows a region with sides of $250 h^{-1} \text{ Mpc}$ and thickness of $75 h^{-1} \text{ Mpc}$ (co-moving with the cosmic expansion). Each simulation contains about two million particles. The upper panels describe a flat low-density model with $\Omega_m = 0.3$ and $\Omega_\Lambda = 0.7$ (L03), and the lower panels show an Einstein–de Sitter model (EdS) with $\Omega_m = 1$. In both cases the amplitude of the power spectrum is consistent with the number density of nearby galaxy clusters^{73,74} and with the large-scale CMB anisotropies⁶⁴. Superimposed on the dark matter distribution, the yellow circles mark the positions of galaxy clusters that would be seen shining in X-rays with a temperature $T > 3 \text{ keV}$, as computed from the cluster mass according to the relation calibrated from hydrodynamical simulations⁴⁵. The size of the circles is proportional to temperature.

an illustrative example of the development of cosmic structures as generated by a modern numerical N -body experiment, simulating the gravitational growth of clustering from given initial conditions and able precisely to resolve individual cluster-sized clumps while following their distribution on scales of several hundreds of Mpc. The mass–luminosity connection makes it possible to compute the correlation function and power spectrum that clusters with a given X-ray luminosity are predicted to have in each model^{31,32,59,60}. The amplitude of the correlation function, and the scales over which clusters are observed to be still inhomogeneously distributed, consistently require a low-density Universe dominated by cold dark matter (CDM), with density parameter $\Omega_m \approx 0.3$. This is shown by the two curves in Figs 2 and 3, chosen to be consistent with CMB anisotropies on small^{15,16} and large⁶⁴ angular scales, but with different Ω_m . Even a small increase of the density parameter from 0.3 to 0.5 produces a clear lack of coherence on scales above $100 h^{-1}$ Mpc. This gives an idea not only of the quality reached by current data, but also of the role future surveys of X-ray clusters could play in high-precision cosmology, when observations at different wavelengths will be combined to pin down the values of cosmological parameters with high accuracy (see ref. 65).

Clusters and the evolution of the Universe

In Fig. 6 we have marked the positions of clusters with a mass corresponding to a temperature larger than 3 keV, which translates into a luminosity of about $10^{44} \text{ erg s}^{-1}$. The two $z > 0$ snapshots correspond to lookback times of 5.7 and 9.0 billion years for the L03 (with $\Omega_m = 0.3$) cosmology and to 6.6 and 9.5 billion years for the Einstein–de-Sitter (EdS) cosmology. Despite the similar pattern produced at the present time ($z = 0$), the past histories of the two models are very different. The most striking feature is probably the fast decay in the abundance of hot, massive clusters as a function of redshift in the $\Omega_m = 1$ model, in contrast to the mild changes visible in the low-density model. This remarkable evolutionary difference represents one of the major motivations for the recent deep X-ray searches of clusters down to fluxes about 1/100 that of the RASS (see refs 66, 67 and references therein). Clusters at $z \approx 0.5$ are no longer considered as exceptions, and even a few examples at $z \geq 1$ are now known⁶⁸. The main result reached by these surveys is the evidence for a weak evolution of the bulk of the cluster population out to $z \approx 1$, again consistent with the picture of a low- Ω_m Universe.

Besides the cluster number density—which basically measures the degree of initial inhomogeneity on scales of less than $10 h^{-1}$ Mpc—the large-scale pattern also evolves in a markedly different way between the two models of Fig. 6, this difference being mainly driven by the value of the density parameter^{69,70} (see Box 1). Could we follow this evolution back in redshift using available X-ray selected clusters? Unfortunately, this is not yet possible: even the largest deep surveys contain only ~ 100 clusters within relatively small patches spread over the whole sky, which gives us no chance to map their distribution within a sufficiently large continuous volume at redshift $z > 0.3$. For this to be possible, a large-area X-ray survey would be needed, reaching a flux limit significantly fainter than the RASS. (For example, detecting a typical cluster with X-ray luminosity $L_X \approx 10^{44} h^{-2} \text{ erg s}^{-1}$ at redshift $z \approx 1$ requires a survey limiting flux about 1/100 of that of the RASS.)

Great progress in X-ray astronomy is foreseen for the next few years, in relation to the newly launched AXAF–Chandra and XMM–Newton X-ray satellites. Also, new, improved X-ray observatories are being considered (see <http://xmm.vilspa.esa.es>). However, none of these missions will be ideal for extended studies of large-scale structure, as they are not designed to perform an all-sky survey, or at least to cover a large enough ($\sim 1,000$ square degrees) contiguous area. Covering the whole sky in a reasonable time and with about 100 times higher sensitivity than the RASS would be feasible even now, given current advances in the technology of X-ray optics⁷¹. An

all-sky survey to this depth would contain more than 100,000 clusters out to $z \approx 1$, with about 15,000 of these above $z \geq 0.5$. Although no such experiment is currently funded, several ideas and proposals are circulating in the scientific community. We can hope that the enormous potential of X-ray clusters to follow the shaping process of the Universe will soon make these projects a reality. $\square$

- Abell, G. O. The distribution of rich clusters of galaxies. *Astrophys. J. Suppl.* **3**, 211–278 (1958).
- Abell, G. O., Corwin, H. G. Jr & Olowin, R. P. A catalog of rich clusters of galaxies. *Astrophys. J. Suppl.* **70**, 1–138 (1989).
- Zwicky, F., Herzog, E., Wild, P., Karpowicz, M. & Kowal, C. *Catalogue of Galaxies and of Clusters of Galaxies* (California Institute of Technology, Pasadena, 1961–68).
- Retzlaff, J., Borgani, S., Gottlöber, S., Klypin, A. & Müller, V. Constraining cosmological models with cluster power spectra. *New Astron.* **3**, 631–646 (1998).
- Miller, C. J. & Batuski, D. J. The power spectrum of rich clusters to scales approaching 1000 Mpc. *Astrophys. J.* (submitted); also preprint astro-ph/0002295 at (xxx.lanl.gov) (2000).
- Zwicky, F. *Helv. Phys. Acta.* **6**, 110 (1933).
- Smith, S. The mass of the Virgo cluster. *Astrophys. J.* **83**, 23–30 (1936).
- Evrad, A. E. The intracluster gas fraction in X-ray clusters—Constraints on the clustered mass density. *Mon. Not. R. Astron. Soc.* **292**, 289–297 (1997).
- Hasinger, G. et al. The ROSAT Deep Survey. I. X-ray sources in the Lockman Field. *Astron. Astrophys.* **329**, 482–494 (1998).
- Ebeling, H. et al. Properties of the X-ray-brightest Abell-type clusters of galaxies (XBACs) from ROSAT All-Sky Survey data—I. The sample. *Mon. Not. R. Astron. Soc.* **281**, 799–829 (1996).
- Ebeling, H. et al. The ROSAT brightest cluster sample—I. The compilation of the sample and the cluster log N-log S distribution. *Mon. Not. R. Astron. Soc.* **301**, 881–914 (1998).
- De Grandi, S. et al. (the REFLEX Team) A flux-limited sample of bright clusters of galaxies from the southern part of the ROSAT All-Sky Survey: The catalog and LOG N-log S. *Astrophys. J.* **514**, 148–163 (1999).
- Böhringer, H. et al. (the REFLEX Team) The ROSAT-ESO flux limited X-ray (REFLEX) galaxy cluster sample. *Astron. Astrophys.* (submitted); also as preprint astro-ph/001226 at (xxx.lanl.gov) (2000).
- Böhringer, H. et al. The northern ROSAT All-Sky (NORAS) galaxy cluster survey I: X-ray properties of clusters detected as extended X-ray sources. *Astrophys. J. Suppl.* **129**, 435–474 (2000).
- de Bernardis, P. et al. (BOOMERANG Collaboration) A flat Universe from high-resolution maps of the cosmic microwave background radiation. *Nature* **404**, 955–959 (2000).
- Balbi, A. et al. (MAXIMA Collaboration) Constraints on cosmological parameters from MAXIMA-1. *Astrophys. J.* **545**, L1–L4 (2000).
- Rood, H. J. Voids. *Annu. Rev. Astron. Astrophys.* **26**, 245–294 (1988).
- Guzzo, L. in *Proceedings of the XIX Texas Symposium on Relativistic Astrophysics & Cosmology* (eds Aubourg, E. et al.) *Nucl. Phys. B (Proc. Suppl.)* **80**, 241 (09/06) (2000); also preprint astro-ph/9911115 at (xxx.lanl.gov) (1999).
- Shectman, S. A. et al. The Las Campanas Redshift Survey. *Astrophys. J.* **470**, 172–188 (1996).
- Peebles, P. J. E. *Principles of Physical Cosmology* (Princeton University Press, Princeton, 1993).
- Tucker, D. L. et al. The Las Campanas redshift survey galaxy–galaxy autocorrelation function. *Mon. Not. R. Astron. Soc.* **285**, L5–L9 (1997).
- Borgani, S. Scaling in the Universe. *Phys. Rep.* **251**, 1–152 (1995).
- Guzzo, L. Is the Universe homogeneous? (on large scales). *New Astron.* **2**, 517–532 (1997).
- Wu, K., Lahav, O. & Rees, M. J. The large-scale smoothness of the Universe. *Nature* **397**, 225–230 (1999).
- Peacock, J. A. *Cosmological Physics* (Cambridge University Press, Cambridge, 1999).
- Coles, P. & Lucchin, F. *Cosmology: The Origin and Evolution of Cosmic Structure* (Wiley, Chichester, 1995).
- Mo, H. J. & White, S. D. M. An analytic model for the spatial clustering of dark matter haloes. *Mon. Not. R. Astron. Soc.* **282**, 347–361 (1996).
- Somerville, R. S. & Primack, J. R. Semi-analytic modelling of galaxy formation: the local Universe. *Mon. Not. R. Astron. Soc.* **310**, 1087–1110 (1999).
- Benson, A. J., Cole, S., Frenk, C. S., Bough, C. M. & Lacey, C. G. The nature of galaxy bias and clustering. *Mon. Not. R. Astron. Soc.* **311**, 793–808 (2000).
- Guzzo, L. et al. (the REFLEX Team) The REFLEX cluster survey: observing strategy and first results on large-scale structure. *The Messenger* **95**, 27–32 (1999).
- Collins, C. A. et al. (the REFLEX Team) The REFLEX galaxy cluster survey II. The spatial correlation function. *Mon. Not. R. Astron. Soc.* **319**, 939–948 (2000).
- Schuecker, P. et al. (the REFLEX Team) The REFLEX galaxy cluster survey II. The spatial correlation function. *Mon. Not. R. Astron. Soc.* (in the press); also preprint astro-ph/0012105 at (xxx.lanl.gov) (2000).
- Bahcall, N. A. & Soneira, R. M. The spatial correlation function of rich clusters of galaxies. *Astrophys. J.* **270**, 20–38 (1983).
- Klypin, A. A. & Kopylov, A. I. The spatial covariance function for rich clusters of galaxies. *Sov. Astron. Lett.* **9**, 41–44 (1983).
- Postman, M. in *Evolution of Large-Scale Structure: from Recombination to Garching* (eds Banday A. J., Sheth, K. & da Costa, L. N.) **270** (Print Partners Ipskamp, Enschede, 1999).
- Kaiser, N. On the spatial correlations of Abell clusters. *Astrophys. J.* **284**, L9–L12 (1984).
- Lumsden, S. L., Nichol, R. C., Collins, C. A. & Guzzo, L. The Edinburgh–Durham southern galaxy catalogue. IV—The cluster catalogue. *Mon. Not. R. Astron. Soc.* **258**, 1–22 (1992).
- Eke, V. R., Cole, S., Frenk, C. S. & Navarro, J. F. Cluster correlation functions in N -body simulations. *Mon. Not. R. Astron. Soc.* **281**, 703–715 (1996).
- Kellogg, E. et al. X-ray observations of the Virgo cluster, NGC 5128, and 3C 273 from the UHURU satellite. *Astrophys. J.* **165**, L49–L52 (1971).
- Gursky, H. et al. A strong X-ray source in the Coma cluster observed by UHURU. *Astrophys. J.* **167**, L81–L84 (1971).
- Giacconi, R. et al. The third UHURU catalog of X-ray sources. *Astrophys. J. Suppl.* **27**, 37–64 (1974).
- Cavaliere, A., Gursky, H. & Tucker, W. H. Extragalactic X-ray sources and associations of galaxies. *Nature* **231**, 437–438 (1971).
- Sarazin, C. *X-ray Emission from Clusters of Galaxies* (Cambridge University Press, Cambridge, 1988).

44. Reiprich, T. H. & Böhringer, H. The empirical X-ray luminosity-gravitational mass relation for clusters of galaxies. *Astron. Nachr.* **320**, 296–299 (1999).
45. Evrard, A. E., Metzler, C. A. & Navarro, J. F. Mass estimates of X-ray clusters. *Astrophys. J.* **469**, 494–507 (1996).
46. Bryan, G. L. & Norman, M. L. Statistical properties of X-ray clusters: Analytic and numerical comparisons. *Astrophys. J.* **495**, 80–99 (1998).
47. White, D. A., Jones, C. & Forman, W. An investigation of cooling flows and general cluster properties from an X-ray image deprojection analysis of 207 clusters of galaxies. *Mon. Not. R. Astron. Soc.* **292**, 419–467 (1997).
48. Arnaud, M. & Evrard, A. E. The $L_X - T$ relation and intracluster gas fractions of X-ray clusters. *Mon. Not. R. Astron. Soc.* **305**, 631–640 (1999).
49. Mushotzky, R. F. & Scharf, C. F. The luminosity-temperature relation at $z = 0.4$ for clusters of galaxies. *Astrophys. J.* **482**, L13–L16 (1997).
50. Della Ceca, R. *et al.* BeppoSAX observations of two high redshift clusters of galaxies: RXJ 0152.7-1357 and MS 2053.7-0449. *Astron. Astrophys.* **353**, 498–506 (2000).
51. Henry, P. Measuring cosmological parameters from the evolution of cluster X-ray temperatures. *Astrophys. J.* **534**, 565–580 (2000).
52. Tozzi, P. & Norman, C. The evolution of X-ray clusters and the entropy of the intra cluster medium. *Astrophys. J.* (in the press); also preprint astro-ph/0003289 at (xxx.lanl.gov) (2000).
53. Bower, R. G. *et al.* The impact of galaxy formation on the X-ray evolution of clusters. *Mon. Not. R. Astron. Soc.* (submitted); also preprint astro-ph/0006109 at (xxx.lanl.gov) (2000).
54. Bartlett, J. G. Sunyaev-Zel'dovich surveys: Analytic treatment of cluster detection. *Astron. Astrophys.* (in the press); also preprint astro-ph/0001267 at (xxx.lanl.gov) (2000).
55. Lahav, O., Edge, A. C., Fabian, A. C. & Putney, A. The spatial distribution of X-ray clusters of galaxies. *Mon. Not. R. Astron. Soc.* **238**, 881–895 (1989).
56. Voges, W. in *Röntgenstrahlung from the Universe* (eds Zimmermann, H. U., Trümper, J. E. & Yorke, H.) 637 (MPE Report No. 263, Max-Planck Institut für Extraterrestrische Physik, Garching bei München, 1996).
57. Romer, A. K. *et al.* The large-scale distribution of X-ray clusters of galaxies. *Nature* **372**, 75–77 (1994).
58. Nichol, R. C., Briel, O. G. & Henry, J. P. The spatial correlation function from an X-ray selected sample of Abell clusters. *Mon. Not. R. Astron. Soc.* **267**, 771–778 (1994).
59. Borgani, S., Plionis, M. & Kolokotronis, V. Cosmological constraints from the clustering properties of the X-ray brightest Abell-type cluster sample. *Mon. Not. R. Astron. Soc.* **305**, 866–874 (1999).
60. Moscardini, L., Matarrese, S., De Grandi, S. & Lucchin, F. The correlation function of X-ray galaxy clusters in the ROSAT All-Sky Survey 1 bright sample. *Mon. Not. R. Astron. Soc.* **314**, 647–656 (2000).
61. Sheth, R. K., Mo, H. J. & Tormen, G. Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes. *Mon. Not. R. Astron. Soc.* (submitted); also preprint astro-ph/9907024 at (xxx.lanl.gov) (1999).
62. Governato, F. *et al.* Properties of galaxy clusters: mass and correlation functions. *Mon. Not. R. Astron. Soc.* **307**, 949–966 (1999).
63. Colberg, J. M. *et al.* Clustering of galaxy clusters in cold dark matter universes. *Mon. Not. R. Astron. Soc.* **319**, 209–214 (2000).
64. Górski, K. M. *et al.* Power spectrum of primordial inhomogeneity determined from the four-year COBE DMR sky maps. *Astrophys. J.* **464**, L11–L14 (1996).
65. Bahcall, N. A., Ostriker, J. P., Perlmutter, S. & Steinhardt, P. J. The cosmic triangle: revealing the state of the Universe. *Science* **284**, 1481–1488 (1999).
66. Rosati, P. *et al.* in *Large Scale Structure in the X-ray Universe* (eds Plionis M. & Georgantopoulos, I.) 13–20 (Atlantis Sciences, Paris, 2000).
67. Gioia, I. in *Proc. IAP2000 Conf. "Constructing the Universe with Clusters of Galaxies"* (eds Gerbal, G. & Durret, E.) (in the press); also preprint astro-ph/0010059 at (xxx.lanl.gov) (2000).
68. Rosati, P. *et al.* An X-ray selected galaxy cluster at $z = 1.26$. *Astron. J.* **118**, 76–85 (1999).
69. Suto, Y., Yamamoto, K., Kitayama, T. & Jing, Y. P. Two-point correlation functions of X-ray-selected clusters of galaxies: Theoretical predictions for flux-limited surveys. *Astrophys. J.* **534**, 551–558 (2000).
70. Moscardini, L., Matarrese, S., Lucchin, F. & Rosati, P. Predicting the clustering of X-ray selected galaxy clusters in flux-limited surveys. *Mon. Not. R. Astron. Soc.* **316**, 283–298 (2000).
71. Burrows, C. J., Burg, R. & Giacconi, R. Optimal grazing incidence optics and its application to wide-field X-ray imaging. *Astrophys. J.* **392**, 760–765 (1992).
72. Eisenstein, D. J. & Hu, P. Baryonic features in the matter transfer function. *Astrophys. J.* **496**, 605–614 (1998).
73. Eke, V. R., Cole, S. & Fenk, C. S. Cluster evolution as a diagnostic for Ω . *Mon. Not. R. Astron. Soc.* **282**, 263–280 (1996).
74. Girardi, M., Borgani, S., Giuricin, G., Mardirossian, F. & Mezzetti, M. The observational mass function of nearby galaxy clusters. *Astrophys. J.* **506**, 45–52 (1998).
75. Viana, P. T. P. & Liddle, A. R. Galaxy clusters at $0.3 < z < 0.4$ and the value of Ω_0 . *Mon. Not. R. Astron. Soc.* **303**, 535–545 (1999).
76. Borgani, S., Rosati, P., Tozzi, P. & Norman, G. Cosmological constraints from the ROSAT deep cluster survey. *Astrophys. J.* **517**, 40–53 (1999).
77. Margon, B. The Sloan digital sky survey. *Phil. Trans. R. Soc. Lond. A* **357**(1750), 93 (1999); also preprint astro-ph/9805314 at (xxx.lanl.gov) (1998).
78. Folkes, S. R. *et al.* (the 2dF team) The 2dF galaxy redshift survey: spectral types and luminosity functions. *Mon. Not. R. Astron. Soc.* **308**, 459–472 (1999).
79. Burles, S. & Tytler, D. The deuterium abundance toward Q1937-1009. *Astrophys. J.* **499**, 699–712 (1998).
80. Lin, H. *et al.* The power spectrum of galaxy clustering in the Las Campanas redshift survey. *Astrophys. J.* **471**, 617–635 (1996).
81. Hoyle, F., Baugh, C. M., Shanks, T. & Ratcliffe, A. The Durham/UKST galaxy redshift survey—VI. Power spectrum analysis of clustering. *Mon. Not. R. Astron. Soc.* **309**, 659–671 (1999).
82. Mellier, Y. Probing the universe with weak lensing. *Annu. Rev. Astron. Astrophys.* **37**, 127 (1999).
83. Borgani, S. *et al.* Measuring Ω_m with the ROSAT Deep Cluster Survey. *Astrophys. J.* (submitted).
84. Carlstrom, J. E. Imaging the Sunyaev-Zel'dovich effect. *Physica Scripta T* **85**, 148–155 (2000).
85. Girardi, M., Giuricin, G., Mardirossian, F., Mezzetti, M. & Boschin, W. Optical mass estimates of galaxy clusters. *Astrophys. J.* **505**, 74–95 (1998).
86. Abadi, M. G., Lambas, D. G. & Muriel, H. Correlation length of X-ray-brightest Abell clusters. *Astrophys. J.* **507**, 526–529 (1998).

Acknowledgements

We thank the REFLEX collaboration for giving us the opportunity to discuss material in advance of publication; C. Collins, F. Governato and M. Urry for reading the manuscript; D. Lazzati and A. Moretti for their help with some of the figures; and M. Girardi for providing the data files on which Fig. 5 is based. We also acknowledge several discussions on X-ray clusters with P. Rosati.

Correspondence and requests for materials should be addressed to S.B. (e-mail: borgani@ts.astro.it) or L.G. (e-mail: guzzo@merate.mi.astro.it).

A scheme for efficient quantum computation with linear optics

E. Knill*, R. Laflamme* & G. J. Milburn†

* Los Alamos National Laboratory, MS B265, Los Alamos, New Mexico 87545, USA

† Centre for Quantum Computer Technology, University of Queensland, St. Lucia, Australia

Quantum computers promise to increase greatly the efficiency of solving problems such as factoring large integers, combinatorial optimization and quantum physics simulation. One of the greatest challenges now is to implement the basic quantum-computational elements in a physical system and to demonstrate that they can be reliably and scalably controlled. One of the earliest proposals for quantum computation is based on implementing a quantum bit with two optical modes containing one photon. The proposal is appealing because of the ease with which photon interference can be observed. Until now, it suffered from the requirement for non-linear couplings between optical modes containing few photons. Here we show that efficient quantum computation is possible using only beam splitters, phase shifters, single photon sources and photo-detectors. Our methods exploit feedback from photo-detectors and are robust against errors from photon loss and detector inefficiency. The basic elements are accessible to experimental investigation with current technology.

Quantum information processing (QIP) uses quantum mechanics for information storage, communication and computation. It enables large improvements in computational efficiency and communication security by exploiting the superposition principle and non-classical correlations of quantum mechanics. Examples include Shor's quantum algorithm for factoring large integers¹, Grover's algorithm for accelerating combinatorial searches² and quantum cryptography for secure communication^{3,4}. Initial concern that quantum coherence may be too fragile to be exploited has been dispelled by theoretical work showing that noise and decoherence are not fundamental obstacles to the implementation of QIP^{5–10}. Consequently, increasing effort is being devoted towards physically realizing quantum computers, and there are many proposals for implementing the necessary quantum devices. Examples of promising technologies include ion traps, quantum dots, Josephson junctions, nuclear spins in silicon and nuclear spins in molecules¹¹.

Quantum effects are particularly easy to observe in optical systems, and it is therefore not surprising that one of the earliest proposals for QIP uses photons to implement quantum logic¹². Optical systems currently constitute the only realistic proposal for long-distance quantum communication and underlie experimental implementations of quantum cryptography^{13–15}. Until now the main obstacle to scalable optical QIP was the apparent need for nonlinear couplings between optical modes. Achieving such couplings at sufficient strengths is possible in principle but is technically difficult¹⁶. As a result, other proposals^{17–19} for using linear optics to benchmark quantum algorithms require exponentially large physical resources.

Here we show the surprising²⁰ result that linear optics is sufficient for efficient QIP with photons. Efficiency in the sense of the theory of computation means with polynomial resources, and we achieve low linear resources. Our proposal for QIP with linear optics requires single photon sources (implementable with active linear optics²¹), beam splitters, phase shifters, photo-detectors, and feedback from photo-detector outputs. A quantum bit (qubit) is realized by one photon in two optical modes (such as horizontal or vertical polarization). Efficient QIP is established by means of three results, each of which constitutes a breakthrough in linear optics QIP. The first result implies that non-deterministic quantum computation²² is possible with linear optics. It is based on a non-linear sign shift between two qubits that uses two additional

photons and post-selection. The sign shift succeeds with probability 1/16, and whether or not it succeeded is known. Although there are no practical applications of non-deterministic quantum computation, it implies that linear optics has features not available to classical deterministic or probabilistic computation. The second result shows that the probability of success of the quantum gates can be increased arbitrarily close to one. The result is based on using entangled states prepared non-deterministically and quantum teleportation^{23,24}. Thus quantum computation is possible in principle with linear optics. The resources needed to make the probability of success close to one with these methods are extremely demanding. The third result shows that with quantum coding, the resources for obtaining accurate encoded qubits are very efficient with respect to the accuracy achieved, thus completing the goal of efficient linear optics quantum computation (LOQC). The coding methods can be adapted to make LOQC fault-tolerant for photon loss, detector inefficiency and phase decoherence. As a result, LOQC can be robustly implemented with resources low enough to suggest practical scalability, making it as promising a technology for QIP as are other proposals.

Bosonic qubits and optical elements

The fundamental units of QIP are qubits, the quantum generalizations of classical bits. A qubit's state space consists of all superpositions $\alpha|0\rangle + \beta|1\rangle$ ($|\alpha|^2 + |\beta|^2 = 1$) of the basic states $|0\rangle$ and $|1\rangle$. A set of qubits can be realized by independent two-state subsystems of a physical system. Bosonic qubits are defined by states of optical modes. An optical mode is a physical system whose state space consists of superpositions of the number states $|n\rangle$, where $n = 0, 1, 2, \dots$ gives the number of photons in the mode. When we consider several qubits or modes, we use labels to distinguish between them. For example, $|20\rangle_{lm}$ (short for $|2\rangle_l|0\rangle_m$) is a state where modes l and m have two and zero photons, respectively. The basic states of a bosonic qubit encoded in modes l_1 and l_2 are $|0\rangle \rightarrow |0\rangle_{l_1}|1\rangle_{l_2}$ and $|1\rangle \rightarrow |1\rangle_{l_1}|0\rangle_{l_2}$. For comprehensive treatments of quantum optics and QIP, see the references^{25–27}.

In addition to instances of an ideal quantum system, a complete implementation of a quantum computer requires a means for state preparation, the ability to apply sufficiently powerful quantum gates, and a readout method. To process information, these elements are combined in quantum networks (see Box 1). The initial state is the vacuum state $|0\rangle$, in which there are no photons in any of

the modes to be used. The basic element that adds photons to the initial state is a single photon source. It can be used to set the state of any given mode to the one-photon state $|1\rangle$. It is sufficient to be able to prepare this state non-deterministically. This means that the state preparation has a non-zero probability of success, and whether or not it succeeded is known.

The simplest optical elements are phase shifters and beam splitters. These elements generate the evolutions implementable by passive linear optics. These evolutions preserve the total photon number, and can be described by their effects on each mode's creation operator, which is defined by $\mathbf{a}^{(b)\dagger}|n\rangle = \sqrt{n+1}|n+1\rangle$. Let U be the unitary operator applied to a state by such an evolution. Using $U|0\rangle = |0\rangle$ gives $U\mathbf{a}^{(b)\dagger}|0\rangle = U\mathbf{a}^{(b)\dagger}U^\dagger U|0\rangle = U\mathbf{a}^{(b)\dagger}U^\dagger|0\rangle = \sum_k u_{kl}\mathbf{a}^{(k)\dagger}|0\rangle$. The coefficients u_{kl} introduced by these equations define a matrix u that must be unitary. Conversely, for every unitary u there is a sequence of phase shifters and beam splitters that implements the corresponding operation up to a global phase²⁸. For a named optical element X , let $u(X)$ be the unitary matrix associated with X according to the above rules. The unitary matrix associated with phase shifter $\mathbf{P}_\theta$ is $u(\mathbf{P}_\theta) = e^{i\theta}$. The unitary matrix associated with beam splitter $\mathbf{B}_{\theta,\phi}$ is

$$u(\mathbf{B}_{\theta,\phi}) = \begin{pmatrix} \cos(\theta) & -e^{i\phi}\sin(\theta) \\ e^{-i\phi}\sin(\theta) & \cos(\theta) \end{pmatrix} \quad (1)$$

We define $\mathbf{B}_\theta = \mathbf{B}_{\theta,0}$.

Phase shifters and beam splitters applied to a bosonic qubit's modes preserve the qubit state space. Their effect can therefore be expressed in the qubit basis using the standard Pauli operators σ_x , σ_y and σ_z . For example, $\mathbf{P}_\theta^{(1)}$ applies $\exp(-i\sigma_z\theta/2)$ up to a global phase shift, and $\mathbf{B}_\theta^{(12)}$ applies $\exp(-i\sigma_y\theta)$. It follows that all one-qubit

rotations can be implemented with linear optics. To achieve the full power of quantum computation we require a two-qubit gate such as the conditional sign flip $c\text{-}z$ defined by $|a\rangle|b\rangle \rightarrow (-1)^{ab}|a\rangle|b\rangle$, where $a, b = 0, 1$ and labels have been omitted.

Readout is accomplished by measuring a mode with a photo-detector, which destructively determines whether one or more photons are present in a mode. We assume that photo-detectors can be applied at any time and that the measurement result can be used to control other optical elements. We need a photon counter, which destructively counts the number of photons in a mode. An approximate photon counter that suffices for our purposes can be designed by using beam splitters and multiple photo-detectors. To measure a mode, we can use beam splitters to distribute the mode's photons evenly over N modes and use a photo-detector on each. The desired count is the number of detectors that 'see' photons. The probability of undercounting given that the photon number is k is at most $k(k-1)/(2N)$. For LOQC, $k \leq 4$.

Nondeterministic conditional sign flip

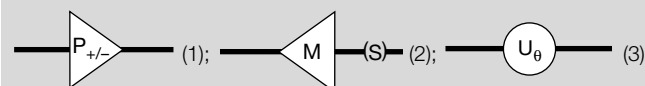
LOQC is based on a series of non-deterministic operations with increasing probability of success. The first operation is a non-deterministic nonlinear sign change on one mode defined by the operation NS: $\alpha_0|0\rangle + \alpha_1|1\rangle + \alpha_2|2\rangle \rightarrow \alpha_0|0\rangle + \alpha_1|1\rangle - \alpha_2|2\rangle$ (with probability 1/4), and can be implemented using the optical network of Fig. 1. Its main features are the use of two ancilla modes with one prepared photon and post-selection based on measuring the ancillas. This procedure can be experimentally verified using techniques similar to those used in a recent Greenberger–Horne–Zeilinger (GHZ) experiment²⁹ (see Supplementary Information). A conditional sign flip $c\text{-}z_{1/16}$ that succeeds with probability 1/16 can be

Box 1

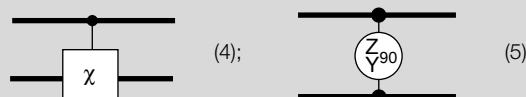
Quantum gates and networks

Quantum information processing (QIP) is accomplished by applying quantum gates and measurements to prepared qubits. The gates evolve the state according to the laws of quantum mechanics. The power of QIP depends on the ability to implement enough evolutions using the available gates. If all unitary evolutions can be approximated up to a global phase, the set of gates is called universal. Standard quantum computation relies on universal gate sets where each gate acts on one or two qubits. One such gate set consists of the one-qubit rotations $U_\phi = \exp(-i\sigma_u\phi/2)$, $U = X, Y$ or Z , where ϕ can be restricted to $\phi = 45^\circ$; and either the conditional sign flip (see text) or one of the 90° rotations $(UV)^{(12)} = \exp(-i\pi\sigma_u^{(1)}\sigma_v^{(2)}/4)$, with $U, V = X, Y$ or Z .

A sequence of state preparations, quantum gates and measurements is called a quantum network. Quantum networks can be depicted by time-space diagrams, with time lines of qubits given by lines running from left to right, and gates by elements that intercept the lines. Our conventions for depicting one qubit gates are:

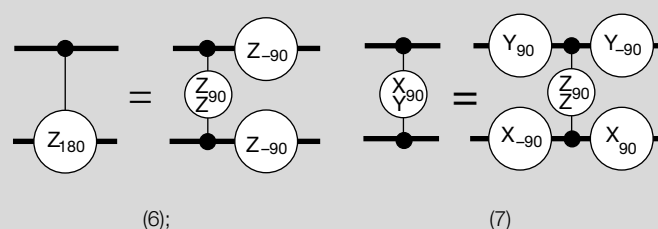


(1) is a preparation gate, with $P = X, Y$ or Z corresponding to preparations of σ_x , σ_y or σ_z eigenstates. For example, if $P_\pm = Z_\pm$, the $|0\rangle$ state is prepared. (2) is a measurement gate, where $M = X, Y$ or Z corresponds to measurements in the eigenbasis of σ_x , σ_y or σ_z . The symbol S denotes the measurement outcome, which can be $+1$ or -1 . (3) is a one-qubit rotation around $U = X, Y$ or Z by angle ϕ (in degrees by default). Two-qubit gates are denoted by



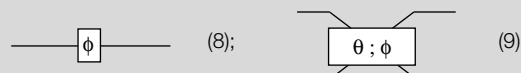
(4) is a conditional sign change by phase χ and applies χ only to the state $|11\rangle$. (5) is a $(ZY)^{(12)}$ rotation.

Many of the gates are equivalent up to one-qubit rotations. Here are some equivalences used in the text:



(7) expresses one gate by conjugating another by $Y_{-90}^{(1)}$ and $X_{90}^{(2)}$.

Optical networks are similar to quantum networks except that the basic systems are optical modes. The basic elements of an optical network drawing are:



(8) shows a phase shifter $\mathbf{P}_\phi$ and (9) a beam splitter $\mathbf{B}_{\theta,\phi}^{(12)}$, where mode 1 is the top mode. If $\phi = 0$, only angle θ may be given in a diagram. State preparation is like (1), with $P_\pm$ replaced by 0 or 1, for the number of photons inserted into the mode. Measurement is like (2), with M replaced by $\mathbf{n}$ and S by R , for the number of photons detected.

implemented with two independent applications of the operation NS as shown in Fig. 2.

Quantum gates by teleportation

Quantum teleportation has proved to be a very versatile tool in QIP^{23,24}. Here we use it to increase the probability of success of coupling gates by reducing the implementation of $c-z$ to a state preparation problem. A basic quantum teleportation protocol T_1 for transferring the state $\alpha_0|0\rangle_1 + \alpha_1|1\rangle_1$ of mode 1 to mode 3 begins by adjoining the ‘entangled’ ancilla state $|t\rangle_{23} = |01\rangle_{23} + |10\rangle_{23}$ (normalization constants omitted). Next, modes 1 and 2 are measured in the basis $|01\rangle_{12} \pm |10\rangle_{12}$, $|00\rangle_{12} \pm |11\rangle_{12}$ (a Bell basis). The measurement consists of two steps. The first determines the parity p of the number of photons in modes 1 and 2 (parity measurement). The second determines the sign s in the superposition. Consider the case where p is odd. If $s = +$, the state of mode 3 is $\alpha_0|0\rangle_3 + \alpha_1|1\rangle_3$. If $s = -$, the state is $\alpha_0|0\rangle_3 - \alpha_1|1\rangle_3$, which can be restored to the initial state by using a phase shifter. For even p , the situation is similar except that $|0\rangle_3$ and $|1\rangle_3$ are flipped (and cannot easily be un-flipped using linear optics). The key property of quantum teleportation is that the input state appears in mode 3 up to a simple transformation without having interacted with mode 3.

The basic teleportation protocol can be implemented in linear optics with success probability $1/2$ by applying a balanced beam splitter to modes 1 and 2 and then measuring the number of photons in the two modes. This partial Bell (or teleportation) measurement (BM_1) determines the parity, and if it is odd, the sign. Using this method, it is possible to implement a conditional sign flip $c-z_{1/4}$ that succeeds with probability $1/4$ (Fig. 3).

To reliably detect photon loss (in the single photon sources, in transmission or by undercounting in detectors), we give another

method, RT_1 , for teleporting a bosonic qubit with success probability $1/2$, which is shown in Fig. 4. The method for obtaining $c-z_{1/4}$ using T_1 works with RT_1 , giving $c-z_{r,1/4}$ with identical failure behaviour (see the caption of Fig. 3). The implementation is in the Supplementary Information.

Near-deterministic quantum teleportation and operations

To improve the probability of successful teleportation to $1 - 1/(n+1)$, we generalize the prepared entanglement by defining $|t_n\rangle_{1\dots(2n)} = \sum_{j=0}^n |1\rangle^j |0\rangle^{n-j} |0\rangle^j |1\rangle^{n-j}$. The notation $|a\rangle^j$ means $|a\rangle|a\rangle\dots|a\rangle$, j times. The modes are labelled from 1 to $2n$, left to right. The state exists in the space of n bosonic qubits, where the k th qubit is encoded in modes $n+k$ and k (in that order). Using the qubit bases, the state $|t_n\rangle$ is $\sum_{j=0}^n |0\rangle^j |1\rangle^{n-j}$. This representation can be used to obtain linear size quantum networks (which are implementable in LOQC) for preparing the state.

A procedure for teleporting the state $\alpha_0|0\rangle_0 + \alpha_1|1\rangle_0$ using $|t_n\rangle$ applies the measurement BM_n , which consists of the $n+1$ point Fourier transform $\hat{F}_{n+1}$ followed by measurement of modes $0\dots n$. $\hat{F}_{n+1}$ is determined by $u(\hat{F}_{n+1})_{kl} = \omega^{kl}/\sqrt{n+1}$, where $\omega = \exp(i2\pi/(n+1))$ and $k, l \in 0\dots n$. It has efficient linear optics implementations^{30,31}.

Suppose BM_n detects k photons altogether. We claim that if $0 < k < n+1$, then the teleported state appears in mode $n+k$ and only needs to be corrected by applying a phase shift. The modes $2n-l$ are in state 1 for $0 \leq l < (n-k)$ and can be reused in future preparations requiring single photons. The modes $2n-l$ are in state 0 for $n-k < l < n$. If $k=0$ we learn that the input state was $|0\rangle_0$ and if $k=n+1$, that it was $|1\rangle_0$. The probability of these two events is $1/(n+1)$, regardless of the input. Both the necessary correction and which mode we teleported to are unknown until after the measurement.

The construction of $c-z_{1/4}$ and $c-z_{r,1/4}$ can be generalized using near-deterministic teleportation. To obtain a conditional sign flip $c-z_{r^2/(n+1)^2}$ that succeeds with probability $n^2/(n+1)^2$, the prepared entanglement consists of two copies of $|t_n\rangle$ modified by applied $c-z$ operations as follows

$$|cs_n\rangle = \sum_{i,j=0}^n (-1)^{(n-i)(n-j)} |1\rangle^i |0\rangle^{n-i} |0\rangle^j |1\rangle^{n-j} |1\rangle^j |0\rangle^{n-j} |0\rangle^i |1\rangle^{n-i} \quad (2)$$

$$= \sum_{i,j=0}^n (-1)^{(n-i)(n-j)} |0\rangle^i |1\rangle^{n-i} |0\rangle^j |1\rangle^{n-j} \quad (3)$$

where the bosonic qubit encoding introduced earlier for $|t_n\rangle$ has been used for the second identity. The teleportation measurements involve the first modes of the two qubits to which $c-z$ is to be

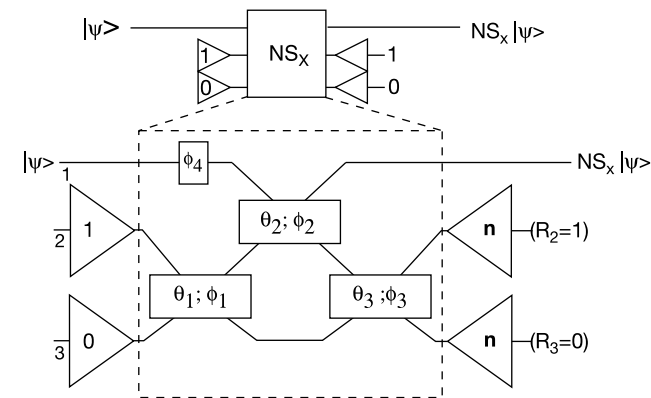


Figure 1 Nonlinear phase shifts on one mode. The numbers at the beginning of the horizontal mode line are their labels. The outlined optical element is abbreviated in future figures by using the top diagram as indicated. The output is accepted only if the photon counters detect one photon in mode 2 and none in mode 3. The subscript x in the top diagram is the phase shift applied and depends on the choice of phases in the optical elements. NS as defined in the text corresponds to $x = -1$ and requires $\theta_1 = 22.5^\circ$, $\phi_1 = 0^\circ$, $\theta_2 = 65.5302^\circ$, $\phi_2 = 0^\circ$, $\theta_3 = -22.5^\circ$, $\phi_3 = 0^\circ$ and $\phi_4 = 180^\circ$. The probability of success is 0.25. Exact expressions for the angles of NS can be determined from the 3×3 unitary matrix u associated with the optical elements²⁸:

$$u = \begin{pmatrix} 1 - 2^{1/2} & 2^{-1/4} & (3/2^{1/2} - 2)^{1/2} \\ 2^{-1/4} & 1/2 & 1/2 - 1/2^{1/2} \\ (3/2^{1/2} - 2)^{1/2} & 1/2 - 1/2^{1/2} & 2^{1/2} - 1/2 \end{pmatrix}.$$

A shift of $|2\rangle_1$ by $x = \exp(i\pi/2)$ is obtained by setting $\theta_1 = 36.53^\circ$, $\phi_1 = 88.24^\circ$, $\theta_2 = 62.25^\circ$, $\phi_2 = -66.52^\circ$, $\theta_3 = -36.53^\circ$, $\phi_3 = -11.25^\circ$ and $\phi_4 = 102.24^\circ$. The probability of success is 0.18082.

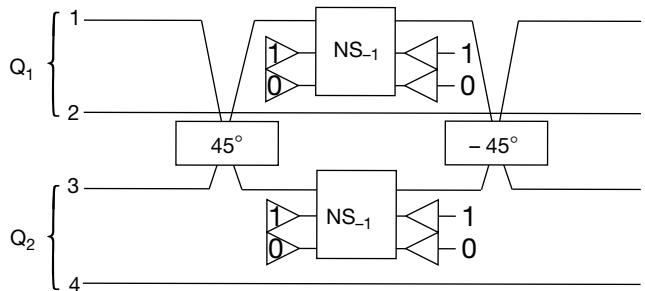


Figure 2 Conditional sign flip implemented with NS operations. Q_1 and Q_2 refer to the bosonic qubits encoded in modes 1, 2, and 3, 4, respectively. Consider the effect of the first beam splitter: When both qubits are in state $|1\rangle$, modes 1 and 3 are in the state $|11\rangle_{13}$, which transforms to $|20\rangle_{13} + |02\rangle_{13}$. In none of the other cases do two photons appear in the same mode. Thus $NS^{(1)}$ and $NS^{(3)}$ have the desired effect. Both of these operations must succeed, so $c-z_{1/16}$ succeeds with probability $1/16$.

applied, and modes $1 \dots n$ and $2n + 1 \dots 2n + n$ (left to right order), respectively. An additional phase correction is needed after the measurement, depending on which modes the output appears in.

To ensure detection of photon loss, the state $|rt_n\rangle$, which generalizes $|rt_1\rangle$, can be used: $|rt_n\rangle = \sum_{j=0}^n |0\rangle^j |1\rangle^{n-j} |0\rangle^{n-j} |1\rangle^j$, written in terms of the qubit encoding. As before, the total number of photons in the modes measured for teleportation is now fixed (at $n + 1$), and any deviation from this results in a detected loss error. The state

needed for the loss-detecting implementation of $c-z$, $c-z_{r,n^2/(n+1)^2}$ is:

$$|rcs_n\rangle = \sum_{i,j=0}^n (-1)^{(n-i)(n-j)} |0\rangle^i |1\rangle^{n-i} |0\rangle^{n-i} |1\rangle^j |0\rangle^{n-j} |1\rangle^j |0\rangle^{n-j} |1\rangle^j \quad (4)$$

The failure-by-measurement behaviour for $c-z_{n^2/(n+1)^2}$ and $c-z_{r,n^2/(n+1)^2}$ can be made the same as that for $c-z_{1/4}$ (see Fig. 3).

Applications of the techniques introduced so far include near-deterministic non-destructive parity measurements, a method for

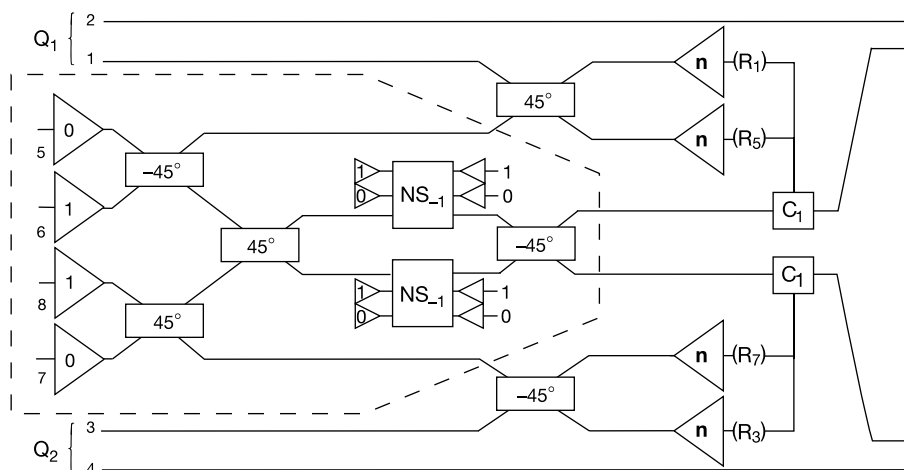


Figure 3 Conditional sign flip with success probability 1/4. The method may be derived as follows²⁴: To implement $c-z$ on two bosonic qubits in modes 1, 2 and 3, 4, respectively, we can teleport the first modes of each qubit to two new modes (labelled 6 and 8) and then apply $c-z$ to the new modes. When using the basic teleportation protocol (T_1), we may need to apply a sign correction. Since this commutes with $c-z$, it is possible to apply $c-z$ to the prepared state before performing the measurements, reducing the implementation of $c-z$ to a state-preparation (outlined) and two teleportations. The two teleportation measurements each succeed with probability 1/2, giving a net success probability of 1/4. The correction operations C_1 consist of applying the phase shifter P_{180° when required by the measurement outcomes. The state preparation needs to be attempted 16 times on average before success, which corresponds to 32 attempted NS operations (without

taking advantage of the ability to avoid an attempt if the first one in a pair failed).

The implementation of $c-z_{1/4}$ fails if one of the two teleportation measurements does not succeed. The following properties hold for failure of $c-z_{1/4}$: (1) the failed teleportation measurements result in an unintentional Z measurement of the corresponding bosonic qubit (2). The teleportation measurements fail independently. (Alternatively, to improve efficiency, one may attempt the measurements sequentially, so as not to perform the second one if the first one fails.) (3) By reintroducing a photon if necessary, the measurements can be assumed to be non-destructive. (4) By applying a phase shifter if necessary, it can be arranged that the effect on the successfully teleported qubit is as if the $c-z$ operation succeeded before the unintentional measurement.

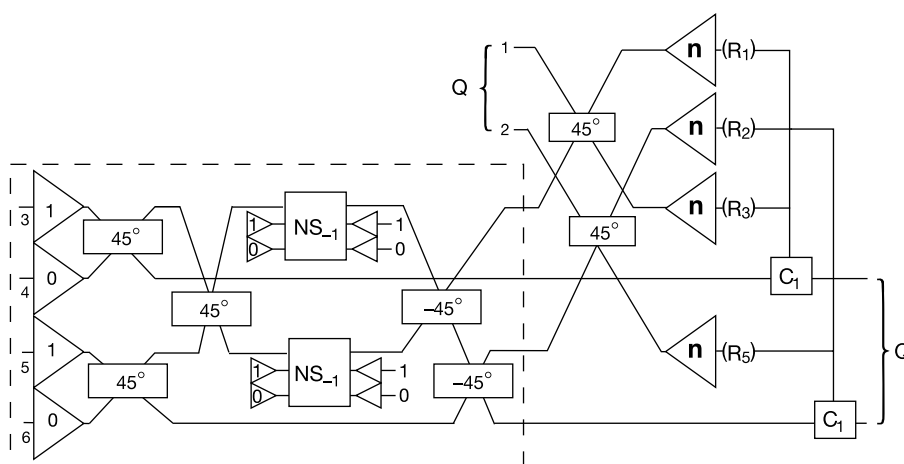


Figure 4 Teleportation with loss detection (RT_1). The outlined box prepares the state $|rt_{3456}\rangle = |01\rangle_{34}|10\rangle_{56} + |10\rangle_{34}|01\rangle_{56}$ using non-deterministic gates. This teleportation protocol has been experimentally tested^{44,45}, using down conversion with post-selection for preparing $|rt_1\rangle$ instead of the preparation network shown above. Given $|rt_1\rangle$, the protocol succeeds with probability 1/2. The pair of NS operations implements a $c-z_{1/16}$ on bosonic qubits encoded in modes 3, 4 and 5, 6, respectively. Thus 32 NS attempts are needed on average before successfully obtaining $|rt_1\rangle$. Without loss, the number of

photons in modes 1, 2, 3, 5 is two. Thus, loss is detected if $R_1 + R_2 + R_3 + R_5 \neq 2$. The teleportation succeeds if $R_1 + R_3 = 1$ and $R_2 + R_5 = 1$, in which case the qubit reappears in modes 4, 6. Failure not due to loss results in a Z measurement of the teleported qubit. Loss of a photon in the incoming qubit or from detector inefficiency is always detected. Assuming no loss in the prepared state or the detectors, RT_1 detects if the input is not a bosonic qubit state (a leakage event) and returns a bosonic qubit. This is necessary for scalable quantum information processing (QIP).

creating entanglement by local measurements of uncorrelated photons shared with beam splitters, and nearly unconditional quantum teleportation and Bell-state measurements with linear optics.

The proof of the claim of this section, the teleportation network for the case $n = 2$, networks for preparing $|cs_2\rangle$ and $|rt_2\rangle$ and

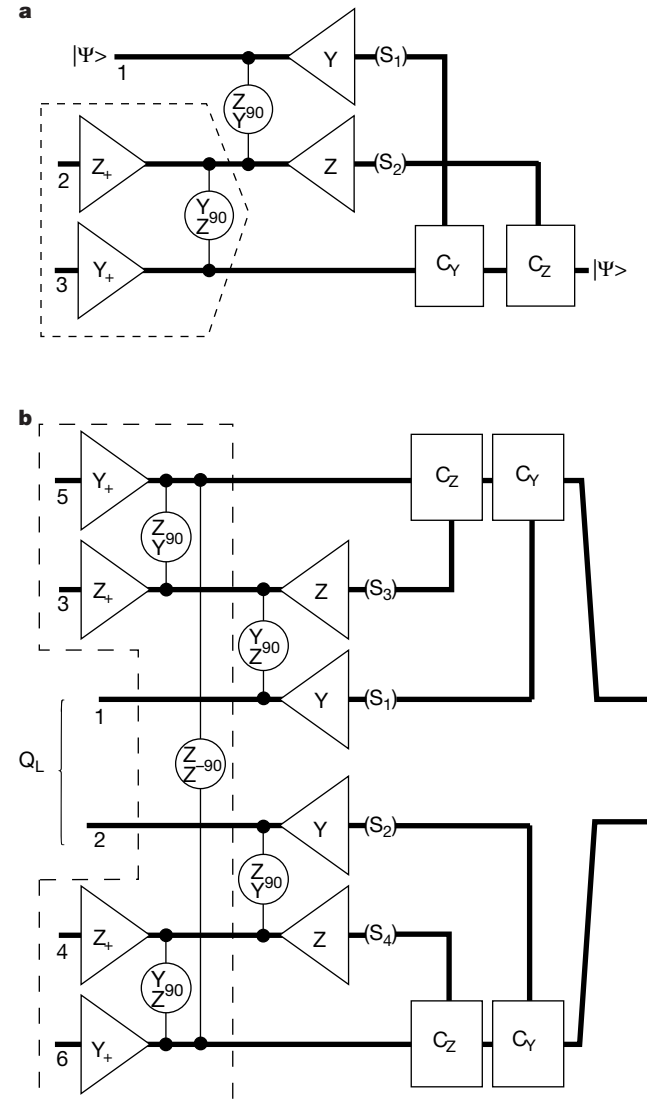


Figure 5 Teleportation networks for the code χ_2 . **a**, Teleportation satisfying that failures of the teleportation step result at worst in a Z -measurement of qubit 1. The networks are based on a variation of the teleportation protocol that exploits the flexibility in the choices for initial states, rotations and measurements to ensure that it behaves well with respect to measurement failures. The correction operations are $C_Z = X_{180^\circ}$ if $S_2 = 1$ and $C_Y = Z_{180^\circ}$ if $S_1 = 1$. The state preparation is outlined and outputs a state denoted by $|tx_2\rangle$. **b**, Teleportation for applying $(Z^{(1)}Z^{(2)})_{90^\circ}$. If $S_3 S_4 = -1$, the phase needs to be corrected with a Z_{180° on both qubits. The prepared state (outlined) is obtained by applying $(ZZ)_{90^\circ}$ to the destination qubits of two copies of $|tx_2\rangle$. The method for applying $(Z^{(1)}Z^{(2)})_{90^\circ}$ is similar, using four copies instead. Both teleportations are attempted. The procedure can only fail with the logical qubit measured in Z . For simplicity, the following failure protocol can be used: If both teleportations fail in any way, we measure the qubits in Z on purpose (if that has not already happened), thus inducing a logical Z measurement. If only one fails, we ensure that the corresponding qubit is measured in Z , then follow the recovery protocol of Fig. 6 using the successfully teleported qubit. With this failure protocol, the logical failure probability for the $Z^{(1)}$ and $Z^{(1)}Z^{(2)}$ rotations is $f_z = (1 - (1 - f)^2)^2 + 2(1 - (1 - f)^2)(1 - f)^2 f_r$, with $f_r = f/(1 - f(1 - f))$ the probability of recovery failure (Fig. 6). Thus $f_z < f$ whenever $f < 1/6.43$.

descriptions of the applications are in the Supplementary Information.

Boosting success with quantum codes

Exponential improvements in the probability of success for gates and state preparation can be obtained by exploiting quantum codes and the failure behaviour of $c - Z_{n^2/(n+1)^2}$. As a result, n need not be large and the difficulty of preparing states such as $|cs_n\rangle$ or $|rcs_n\rangle$ is lessened. We give a method based on a two-qubit code, χ_2 . This method can be used to define logical qubits with greatly improved success probabilities and robustness, provided that the given qubits are sufficiently controllable. As a result it is possible to iterate the method to efficiently achieve essentially perfect QIP. This iteration is known as concatenation and underlies the accuracy-threshold theorems of fault-tolerant quantum computation^{6–9}.

From now on, we use qubit based quantum networks and rely on the following list of operations implementable in LOQC with bosonic qubits according to the techniques of the previous sections: (1) X , Y and Z eigenstate (eigenvalue $+1$ or -1) preparation; (2) X , Y and Z measurements; (3) X_{180° , Y_{180° and Z_{180° rotations; (4) X_ϕ rotations; (5) Z_{90° rotation; (6) $(Z^{(1)}Z^{(2)})_{90^\circ}$ rotation. For the moment we assume that the optical elements, single-photon sources and photon counters are error-free. Operations (1) to (4) always succeed. The $(Z^{(1)}Z^{(2)})_{90^\circ}$ rotation fails independently on qubits 1 and 2 with probability f . If $c - Z_{n^2/(n+1)^2}$ is used, then $f = 1/(n + 1)$. The Z_{90° rotation always succeeds in LOQC, although after the first encoding it fails with probability f . A qubit on which an operation fails is measured in Z after the rotation has been applied. The Y_{90° , $(YZ)_{90^\circ}$ and $(YY)_{90^\circ}$ rotations can be implemented by conjugation of Z_{90° or $(ZZ)_{90^\circ}$ with failure-free X rotations. The failure mode of these rotations is similar to that for the $(ZZ)_{90^\circ}$ rotation, with commuting Y measurements replacing Z measurements.

To encode a qubit we define its logical states $|0\rangle_L$ and $|1\rangle_L$ by $|0\rangle_L = |00\rangle + |11\rangle$ and $|1\rangle_L = |01\rangle + |10\rangle$. This is an instance of a stabilizer code^{32–35}. In this context it is convenient to use the abbreviation $U = \sigma_u$ for $U = X, Y, Z$. With encoding qubits labelled 1, 2, the logical X , Y and Z operators are given by $X^{(L)} = X^{(1)} = X^{(2)}$, $Z^{(L)} = Z^{(1)}Z^{(2)} =_L - Y^{(1)}Y^{(2)}$ and $Y^{(L)} = Y^{(1)}Z^{(2)} =_L Z^{(1)}Y^{(2)}$, where we introduced the notation $=_L$ to denote identity when restricted to the code space spanned by $|0\rangle_L$, $|1\rangle_L$. To destructively measure one of the logical operators, it suffices to measure each qubit; it is straightforward to obtain nondeterministic state preparation networks (see the Supplementary Information). Any rotation $X_\phi^{(L)}$ can be implemented by applying $X_\phi^{(1)}$ or $X_\phi^{(2)}$. The 180° logical rotations can be applied by using the corresponding 180° rotations directly on the qubits, a feature satisfied by all stabilizer codes³⁶. For example, to apply $Y_{180^\circ}^{(L)}$ apply both $Y_{180^\circ}^{(1)}$ and $Z_{180^\circ}^{(2)}$.

The logical operations introduced so far can be done without failure. To implement the $Z_{90^\circ}^{(L)}$ and $(ZZ)_{90^\circ}^{(L)}$ rotations with failure probabilities much less than f requires the teleportation networks shown in Fig. 5, which have the property that at worst, the teleported qubit is measured in Z . As described in the captions of Figs 5 and 6, the failure probability f_z of these logical rotations satisfies $f_z < O(f^2)$ and $f_z < f$ whenever $f < 1/6.43$.

The methods can be improved in three ways: first, by better exploiting the flexibility in state preparation and responses to failures; second, by using classical linear codes like the repetition codes; and third, by encoding more than one qubit into one block. With these techniques it is possible to achieve $f_z < f$ for $f < 1/2$ (see Supplementary Information).

Scalability and resource requirements

A scalable information processing system requires that one can deal with errors that occur in the physical implementation. For LOQC, dominant sources of errors are photon loss (at the single photon source or during processing), detector inefficiency (which can be

viewed as photon loss) and phase errors. Photon loss can be dealt with by using the loss-detecting implementations of $c-z$. The probability f_l of loss for an LOQC operation can be predicted from the characteristics of the optical devices. The possibility of loss introduces a new failure mode, where nothing is known about what happened to the state of the qubit. This is the erasure model of errors³⁷. A good implementation of LOQC ensures that $f_l \ll f$, so that we can first improve f using the techniques already discussed, and then deal with the problem of erasures. Compensating for erasures is much easier than dealing with general errors, with pessimistic estimates of $f_l \lesssim 0.01$ (ref. 38) for quadratic improvements. Unlike photon loss, phase errors are not detected by the networks discussed so far. Happily, phase-error correction can be integrated into the methods for reducing f using codes that generalize χ_2 based on classical repetition codes. These codes can correct unknown phase errors in up to half the qubits. More details on erasure and phase error correcting codes are in the Supplementary Information.

The methods introduced so far suffice for implementing accurate quantum gates on logical qubits in the presence of intrinsic failures of LOQC, and sufficiently low photon loss and phase errors. Scalable quantum computation is possible provided that any remaining errors in the logical operations fall below a threshold. There is evidence that the relevant threshold may be above 0.0001 (D. Gottesman and J. Preskill, unpublished work). Achieving such low error is experimentally challenging for any device, although optimism is justified by the observations that many of the errors are due to improper calibration of classical control parameters, and these are often controllable well below the estimated threshold. An example is pulse phase in nuclear magnetic resonance. Another reason for optimism is that at least for quantum communication, the threshold is well above 0.01 (ref. 39). As all viable proposals for long-distance quantum communication are based on optics, this may be the first scalable application of LOQC.

Resources contributing toward a logical quantum gate based on LOQC can be counted in two ways: as total and as conditional resources. The total resources are given by the number of optical operations required on average. This depends on the success probabilities of the component state preparations and the desired success probability for the logical operations. As most of the resources are used in independent state preparation steps, an implementation of LOQC can be based on massively parallel state

factories. It is thus natural to consider the conditional resources, which are the number of optical operations that successfully contribute toward a logical quantum gate. Their significance is that the error of an operation conditional on success can be estimated by multiplying the conditional error of the optical elements by the conditional resource count. Detailed resource analyses are yet to be done. However, it can be shown that failure probabilities below 5% can be achieved using only two iterations of χ_2 , requiring about 300 successful $c-z_{9/16}$ operations per logical two-qubit gate³⁸.

An implementation of LOQC requires careful mode matching, rapidly controllable delay lines or good synchronization of pulses, tunable beam splitters and phase shifters, single photon sources and high-efficiency fast photo-detectors for single photon detection. Speed is needed to be able to select successful state preparations before photon loss becomes too large. Tunable optical elements can be made using polarizers and polarizing beam splitters. Non-deterministic single photon sources can be constructed with parametric down converters²¹, although a better method is to use one of the schemes for single photon sources that have recently been proposed^{40,41}. The best photon counters currently have efficiencies of about 0.9 at optical frequencies⁴². This is sufficient for experimentally implementing the basic elements of LOQC. Higher efficiencies are required for implementing the more complex teleportation and quantum gate operations with sufficiently low error conditional on success.

The preliminary resource counts discussed above imply large but not excessive resource overheads per reliable quantum gate. The need for robustness requires non-trivial resource overheads in all implementations of QIP, so this suggests that scalable quantum computation using LOQC is comparable in complexity to other proposals. LOQC has the advantage in several respects. In particular, there is no need for low temperature for the basic optical elements (except perhaps in the single photon sources and the photo-detectors, depending on implementation), and photons naturally maintain their coherence over timescales that are long compared to the basic control operations. Furthermore, the sources of noise are better understood and do not depend on difficult-to-predict or difficult-to-measure thermal interactions. However, all proposals until now, including LOQC, require that various technologies can be made to work together to obtain high fidelities in operations.

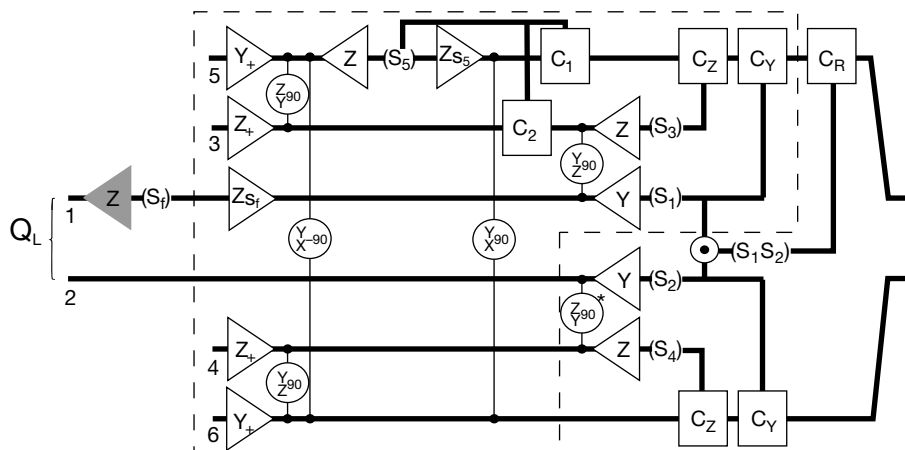


Figure 6 Recovery from Z measurement. A state is prepared using two instances of $|tx_2\rangle$ modified by projecting with $I + XX$. The gates C_1 and C_2 correct for measuring $S_5 = -1$ by applying $C_1 = Z_{180^\circ}$ to qubit 5 and $C_2 = X_{180^\circ}$ to qubit 3. The two teleportations are then attempted. The network assumes that we need to recover from a Z measurement of qubit 1 (shown in grey). In this case, qubit 1 can be absorbed into a state preparation; which one depends on the measurement outcome. This avoids being affected by failures in the top teleportation. The parts of the network which can be performed in a

non-deterministic state preparation are outlined. An XX measurement of the teleported qubits becomes recorded in the teleportation measurements. The recovered state is obtained by applying $C_R = Z_{180^\circ}$ if $S_1 S_2 = -1$. If the pre-measurement coupling gate marked by an asterisk fails with a Y -measurement only, then we can retry the recovery process using a new prepared state. The probability of this failure event is $f(1-f)$, so the total failure probability f_r of recovery satisfies $f_r = f + f(1-f)f_r$, whence $f_r = f/(1-f(1-f))$.

Discussion

Linear optics was believed to be insufficient for quantum computation because every implementable evolution can be understood in terms of a small unitary matrix, contrary to expectations of exponential complexity. Furthermore, passive linear optics does not involve particle interactions other than those imposed by statistics and can be understood in terms of classical wave mechanics. There is, however, a hidden non-linearity in LOQC (in the photo-detectors) and our techniques effectively transfer this non-linearity to the bosonic qubits, thus enabling universal quantum computation.

There are other options for implementing LOQC. Particularly interesting is an idea⁴³ that involves encoding qubits in the phase space of a mode. Universal computation in this system requires active linear optics and a nonlinearly prepared state, but has the advantage of being intrinsically robust against errors involving shifts in the canonically conjugate variables. It may be possible to combine approaches to achieve robust and efficient LOQC even more easily. □

Received 24 July; accepted 13 November 2000.

- Shor, P. W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM J. Comput.* **26**, 1484–1509 (1997).
- Grover, L. K. Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.* **79**, 325–328 (1997).
- Wiesner, S. Conjugate coding. (*Original Manuscript ~ 1969*) *Sigact News* **15**, 78–88 (1983).
- Bennett, C., Brassard, F., Brassard, G., Salvail, L. & Smolin, J. Experimental quantum cryptography. *J. Cryptol.* **5**, 3–28 (1992).
- Shor, P. W. in *Proceedings of the 37th Symposium on the Foundations of Computer Science (FOCS)* 56–65 (IEEE Press, Los Alamitos, 1996).
- Aharonov, D. & Ben-Or, M. in *Proceedings of the 29th Annual ACM Symposium on the Theory of Computation (STOC)* 176–188 (ACM Press, New York, 1996).
- Kitaev, A. Y. Quantum computations: algorithms and error correction. *Russian Math. Surv.* **52**, 1191–1249 (1997).
- Knill, E., Laflamme, R. & Zurek, W. H. Resilient quantum computation. *Science* **279**, 342–345 (1998).
- Preskill, J. Reliable quantum computers. *Proc. R. Soc. Lond. A* **454**, 385–410 (1998).
- Steane, A. Efficient fault-tolerant quantum computing. *Nature* **399**, 124–126 (1999).
- Experimental proposals for quantum computation. (Special focus issue) *Fort. Phys.* **48**, 767–1138 (2000).
- Milburn, G. J. Quantum optical Fredkin gate. *Phys. Rev. Lett.* **62**, 2124–2127 (1988).
- Hughes, R. J., Morgan, G. L. & Peterson, C. G. Quantum key distribution over a 48 km optical fibre network. *J. Mod. Optics* **47**, 533–547 (2000).
- Tittle, W., Brendel, J., Gisin, N. & Zbinden, H. Long-distance Bell-type tests using energy-time entangled photons. *Phys. Rev. A* **59**, 4150–4163 (1999).
- Townsend, P., Rarity, J. & Tapster, P. Single photon interference in 10 km long optical fibre interferometer. *Electron. Lett.* **29**, 1291–1293 (1993).
- Turchette, Q. A., Hood, C. J., Lange, W., Mabuchi, H. & Kimble, H. J. Measurement of conditional phase shifts for quantum logic. *Phys. Rev. Lett.* **74**, 4710–4713 (1995).
- Cerf, N. J., Adami, C. & Kwiat, P. G. Optical simulation of quantum logic. *Phys. Rev. A* **57**, R1477–R1480 (1998).
- Howell, J. C. & Yeazell, J. A. Reducing the complexity of linear optics quantum circuits. *Phys. Rev. A* **61**, 052303/1–5 (2000).
- Kwiat, P. G., Mitchell, J. R., Schwindt, P. D. D. & White, A. G. Grover's search algorithm: An optical approach. *J. Mod. Optics* **47**, 257–266 (2000).

- Lütkenhaus, N., Calsamiglia, J. & Suominen, K.-A. Bell measurements for teleportation. *Phys. Rev. A* **59**, 3295–3300 (1999).
- Hong, C. K. & Mandel, L. Experimental realization of a localized one-photon state. *Phys. Rev. Lett.* **56**, 58–60 (1986).
- Adleman, L. M., DeMarrais, U. & Huang, M.-D. A. Quantum computability. *SIAM J. Comput.* **26**, 1524–1540 (1997).
- Bennett, C. H. et al. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **70**, 1895–1899 (1993).
- Gottesman, D. & Chuang, I. L. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature* **402**, 390–393 (1999).
- Walls, D. F. & Milburn, G. J. *Quantum Optics* (Springer, Berlin, 1994).
- Aharonov, D. in *Annual Reviews of Computational Physics VI* (ed. Stauffer, D.) (World Scientific, Singapore, 1999).
- DiVincenzo, D. The physical implementation of quantum computation. *Fort. Phys.* **48**, 771–793 (2000).
- Reck, M., Zeilinger, A., Bernstein, H. J. & Bertani, P. Experimental realization of an discrete unitary operator. *Phys. Rev. Lett.* **73**, 58–61 (1994).
- Bouwmeester, D., Pan, J.-W., Daniell, M., Weinfurter, H. & Zeilinger, A. Observation of three-photon Greenberger-Horne-Zeilinger entanglement. *Phys. Rev. Lett.* **82**, 1345–1349 (1999).
- Wehls, G., Reck, M., Weinfurter, H. & Zeilinger, A. All-fiber three-path Mach-Zehnder interferometer. *Opt. Lett.* **21**, 302–304 (1996).
- Cormen, T. H., Leiserson, C. E. & Rivest, R. L. *Introduction to Algorithms* 795 (MIT Press, Cambridge, MA, 1990).
- Shor, P. W. Scheme for reducing decoherence in quantum computer memory. *Phys. Rev. A* **52**, 2493–2496 (1995).
- Steane, A. Multiple particle interference and quantum error correction. *Proc. R. Soc. Lond. A* **452**, 2551–2577 (1996).
- Calderbank, A., Rains, E., Shor, P. & Sloane, N. Quantum error correction and orthogonal geometry. *Phys. Rev. A* **78**, 405–408 (1997).
- Gottesman, D. A class of quantum error-correcting codes saturating the quantum hamming bound. *Phys. Rev. A* **54**, 1862–1868 (1996).
- Gottesman, D. A theory of fault-tolerant quantum computation. *Phys. Rev. A* **57**, 127–137 (1998).
- Grassl, M., Beth, T. & Pellizzari, T. Codes for the quantum erasure channel. *Phys. Rev. A* **56**, 33–38 (1997).
- Knill, E., Laflamme, R. & Milburn, G. Thresholds for linear optics quantum computation. Preprint quant-ph/0006120 at (xxx.lanl.gov) (2000).
- Dür, W., Briegel, H.-J., Cirac, J. I. & Zoller, P. Quantum repeaters based on entanglement purification. *Phys. Rev. A* **59**, 169–181 (1999).
- Kim, J., Benson, O., Kan, H. & Yamamoto, Y. A single-photon turnstile device. *Nature* **397**, 500–503 (1999).
- Foden, C. L., Talyanskii, V. I., Milburn, G. J., Leadbeater, M. L. & Pepper, M. High frequency acousto-electric single photon source. *Phys. Rev. A* **62**, 011803(R)/1–4 (2000).
- Takeuchi, S., Yamamoto, Y. & Hogue, H. H. Development of a high-quantum-efficiency single-photon counting system. *Appl. Phys. Lett.* **74**, 1063–1065 (1999).
- Gottesman, D., Kitaev, A. & Preskill, J. Encoding a qubit in an oscillator. Preprint quant-ph/0008040 at (xxx.lanl.gov) (2000).
- Bouwmeester, D., Pan, J., Mattle, K., Eibl, M., Weinfurter, H. & Zeilinger, A. Experimental quantum teleportation. *Nature* **390**, 575–579 (1997).
- Boschi, D., Branca, S., Martini, F. D., Hardy, L. & Popescu, S. Experimental realization of teleporting an unknown pure quantum state via dual classical and Einstein-Podolski-Rosen channels. *Phys. Rev. Lett.* **80**, 1121–1125 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank P. Kwiat and A. White for help and discussions.

Correspondence and requests for materials should be sent to E. Knill (e-mail: knill@lanl.gov).

Dual origin of tribosphenic mammals

Zhe-Xi Luo*, Richard L. Cifelli† & Zofia Kielan-Jaworowska‡

*Section of Vertebrate Paleontology, Carnegie Museum of Natural History, Pittsburgh, Pennsylvania 15213, USA

†Oklahoma Museum of Natural History, 2401 Chautauqua, Norman, Oklahoma 73072, USA

‡Institute of Paleobiology, Polish Academy of Sciences, ulica Twarda 51/55, PL-00-818 Warszawa, Poland

Marsupials, placentals and their close therian relatives possess complex (tribosphenic) molars that are capable of versatile occlusal functions. This functional complex is widely thought to be a key to the early diversification and evolutionary success of extant therians and their close relatives (tribosphenidans). Long thought to have arisen on northern continents, tribosphenic mammals have recently been reported from southern landmasses. The great age and advanced morphology of these new mammals has led to the alternative suggestion of a Gondwanan origin for the group. Implicit in both biogeographic hypotheses is the assumption that tribosphenic molars evolved only once in mammalian evolutionary history. Phylogenetic and morphometric analyses including these newly discovered taxa suggest a different interpretation: that mammals with tribosphenic molars are not monophyletic. Tribosphenic molars evolved independently in two ancient (holotherian) mammalian groups with different geographic distributions during the Jurassic/Early Cretaceous: an australosphenidan clade endemic to Gondwanan landmasses, survived by extant monotremes; and a boreosphenidan clade of Laurasian continents, including extant marsupials, placentals and their relatives.

Because most mammals are only represented in the fossil record by their teeth, dental evidence has prominently figured in interpreting the relationships of early mammals. The highly distinctive tribosphenic molars, capable of both shearing (*sphen*) and grinding (*tribein*) occlusal functions, are the most important dental feature of marsupials and placentals (extant therians) and their close fossil relatives, collectively known as tribosphenidans¹. The tribosphenic molar has been used to distinguish these therians from other Mesozoic mammals^{1–4}, including stem holotherians such as “symmetrodonts” and “eupantotheres”^{3,4}, and from living monotremes.

The traditional view that tribosphenic molars arose in the Early Cretaceous, and that the early diversification of therians took place on northern continents⁵, has been challenged by recent discoveries on Gondwanan landmasses. *Ambondro*, from Madagascar, has fully tribosphenic molars⁶ yet is of Middle Jurassic age, and is therefore some 25 Myr older than the earliest northern therians. *Ausktribosphenos*, a tribosphenic mammal from the Early Cretaceous of Australia, is thought to have been a placental, possibly even a member of the living hedgehog family^{7,8}. Thus, these new fossils have important implications for reconstructing the paleobiogeography and timing of divergence of the main extant mammalian groups. Here we re-evaluate the phylogenetic relationships of these newly discovered mammals, with respect to other Mesozoic and extant mammal groups.

Phylogenetic relationships

Parsimony analysis of the available evidence from the dentition and mandible (Fig. 1; also see Methods) places the tribosphenic mammals *Ambondro*⁶ and *Ausktribosphenos*^{7,8}, together with the earliest known monotreme *Steropodon*⁶ (Australia, Early Cretaceous), in a monophyletic group, which we term the Australosphenida (concept modified from Ausktribosphenida⁷; see ‘Systematic Palaeontology’). This group is characterized by a well-developed, continuous mesial cingulid that wraps around and extends to the lingual side of the trigonid (Fig. 2); a mesiodistally short, buccolingually broad talonid; and reduced height of the trigonid—at least when compared with other Jurassic/Early Cretaceous holotherians. Notably, *Ausktribosphenos* retains plesiomorphies of the jaw as seen in stem mammals⁹, and we provisionally recognize some of these primitive characters to be preserved also in *Steropodon*. By contrast, the Boreosphenida, a clade of northern mammals with tribosphenic molars^{10–12} (Fig. 1), have distinctive cingulid cuspules (cuspule e, cuspule f, or both) but lack a continuous mesial cingulid. The mesial

cingulid does not wrap around or extend to the lingual side of the molars. Furthermore, boreosphenidans and proximal relatives, such as *Henkelotherium*, are more derived in that they have a mandibular angle far more posteriorly positioned than australosphenidans, *Kuehneotherium*, *Morganucodontids* and *Sinoconodon*.

Further dental apomorphies are consistent with the monophyly of australosphenidans. For example, *Steropodon*¹³ and *Ausktribosphenos*⁷ both have the ‘twinned’ paraconid and metacoenid, and a well-developed lingual cusp on the talonid (Fig. 2). *Ambondro*⁶ and *Ausktribosphenos*^{7,8} share an unusual feature in which the premolars seem to be molariform, with a distinctive triangulation of the three main cusps and a prominent lingual cingulid. The posterior part of the premolar is buccolingually wide. This derived feature is shared by *Obdurodon dicksoni*¹⁴ (Australia, Miocene), an ornithorhynchid monotreme^{14,15} widely accepted to be a relative of *Steropodon* (in which the premolar is not preserved¹³). None of these derived premolar features of fossil ornithorhynchids, *Ambondro* and *Ausktribosphenos* is present in the earliest known eutherians^{10,11}, metatherians¹², nor stem taxa of the northern tribosphenidans^{16–18}. They are also absent in a wide variety of non-tribosphenic (eupantothere) mammals including *Peramus*¹⁷ and *Henkelotherium*¹⁸.

Australosphenidans may have affinities with the northern holothere *Shuotherium*^{19–21} (Fig. 1), some unknown early “symmetrodonts”, or the holotherian Dryolestidae¹⁴. However, by any previous interpretation they are independent from boreosphenidans and their proximal relatives, including *Henkelotherium*¹⁹ and pre-tribosphenic *Peramus*^{17,22}. The clade of *Ausktribosphenos*, *Ambondro* and the monotreme *Steropodon* (Fig. 1), as inferred from mandibular and postcanine tooth morphology, is consistent with the phylogenies established by independent studies of basicranial^{23,24} and postcranial^{25–27} features of the main Mesozoic mammal groups documented by well-preserved fossils. In our expanded analysis (Fig. 1b), we added the cranial and postcranial features for those better-preserved taxa (for example, *Morganucodon*, *Jeholodens* and *Henkelotherium*). The simultaneous analysis (Fig. 1b) of cranial and postcranial data, together with the dental and mandibular characters, corroborates the australosphenidan clade and the boreosphenidan clade. Hence, the hypothesis of a diphyletic origin for tribosphenic mammals is supported by morphological characteristics other than teeth and mandibles.

Independent multivariate analysis of the shape of the lower molar shows that *Ausktribosphenos* and *Steropodon* are grouped together

(Fig. 3; *Ambondro* is not sufficiently complete for inclusion). These australosphenidans occupy their own morphospace and are distinct from the stem taxa of boreosphenidans, metatherians and eutherians (Fig. 3). Both parsimony analysis of discrete characters (Fig. 1) and shape analysis (Fig. 3) suggest that *Ausktribosphenos* does not belong to the crown group of therians^{9,15}, and that it may have close affinities to the monotreme *Steropodon*^{15,28}.

Extant monotremes are considered to represent an archaic clade, although they are also highly specialized in their own ways^{29,30}. The Early Cretaceous monotreme *Steropodon*¹³ bears some dental similarities to *Peramus*²² and also to dryolestoids¹⁴, which leads to the hypothesis that monotremes may be related to some pre-tribosphenic therian mammals. All recent morphological parsimony analyses^{2,23–27} of the Mesozoic mammals with well-preserved fossils, however, have consistently placed ornithorhynchid monotremes (members of the australosphenidan clade) at a more basal position on the mammalian tree than the pretribosphenic, tribosphenic and extant therians. The australosphenidan clade that includes the monotreme *Steropodon* is consistent with these recent analyses of the combined dental, skull and skeletal characteristics^{23–27}. Our

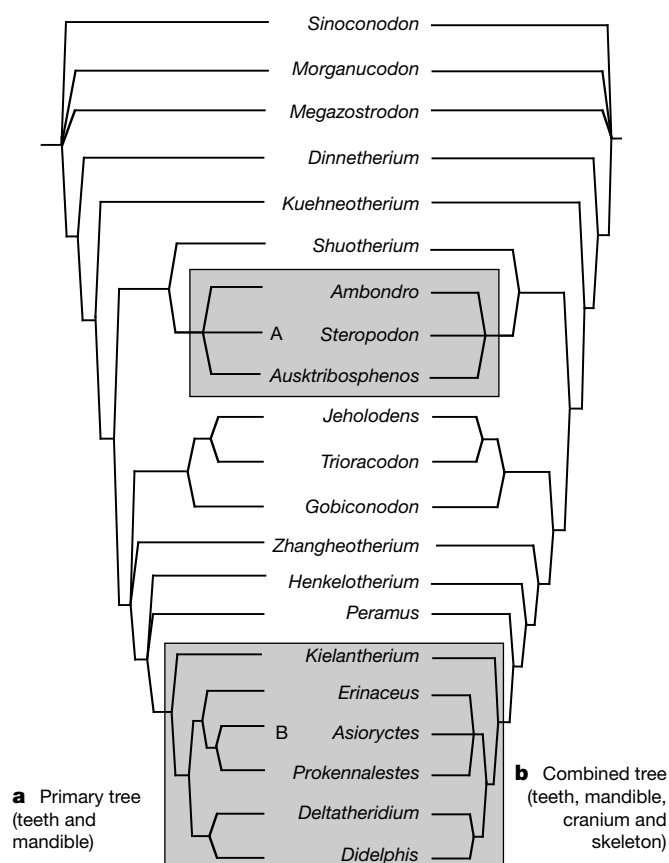


Figure 1 Phylogenetic relationships of main taxa of therian tribosphenic mammals. The Australosphenida (Box A) and Boreosphenida (Box B) are both monophyletic and separate from each other. **a**, Primary analysis: the strict consensus of 18 equally parsimonious and shortest trees from the 55 dental and mandibular characters that are known for the southern tribosphenic mammals (see Supplementary Information). **b**, Extended analysis: the strict consensus of eight equally parsimonious and shortest trees from an expanded analysis including cranial and postcranial data for the better preserved taxa (118 informative characters; see Supplementary Information). Placement of the australosphenidan clade on the mammalian tree is consistent, whether it is based on just the dental and mandibular features known for southern taxa (**a**), or on combined data of the dentition, mandible, plus cranium^{23,24} and postcranium^{25–27} (**b**). Apomorphies supporting the main cladogram nodes are presented in Supplementary Information (see Methods for details).

phylogeny is also consistent with the latest molecular evidence (on the basis of imprinted genes³¹) that would place monotremes outside the extant therians; but is contrary to the “Marsupionta” hypothesis³² that would link monotremes to extant marsupials, exclusive of placentals (on the basis of mitochondrial genes³³).

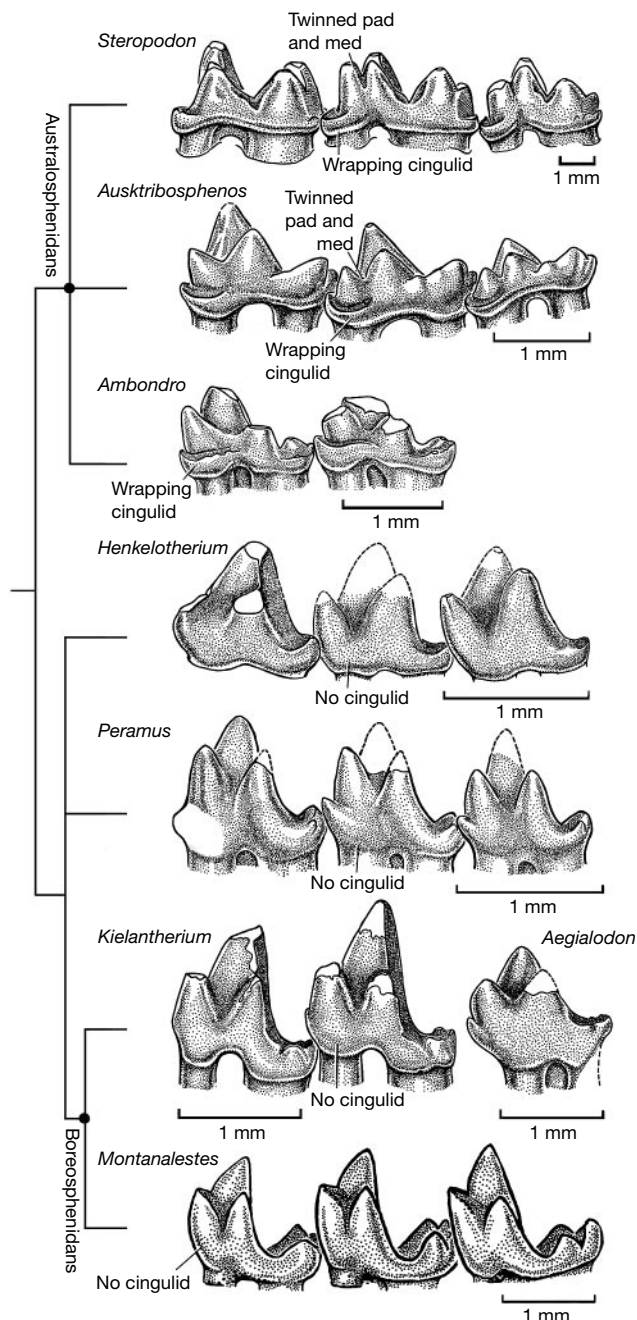


Figure 2 Comparison of molars of australosphenidans, pretribosphenids ('eupantotheres': *Henkelotherium* and *Peramus*) and boreosphenidans, in lingual view. Tree topology pruned from Fig. 1a. Lower molars of australosphenidans are characterized by a mesial cingulid wrapping around to the lingual side of the paraconid (pad), and the reduced height of the trigonid relative to the talonid. *Steropodon* and *Ausktribosphenos* share the derived feature of the twinned paraconid and metaconid (med) and the hypertrophied lingual cusp of the talonid. *Ambondro* and *Ausktribosphenos* share the derived features of a lingual cingulid on molariform premolars (not illustrated). None of these derived dental characters is present in the northern eupantotheres (*Henkelotherium*, *Peramus*), primitive tribosphenidans (*Kielantherium*, *Aegialodon*) or therians (*Montanalestes*).

Molar evolution

The tribosphenic lower molar is defined by a basin-like heel (talonid), which grinds (*tribein*) with the large inner cusp (protocone) on the upper molar—functionally analogous to mortar-to-pestle grinding. This is in addition to a wedge-like trigonid (*sphen*) to shear with the crests of the corresponding upper tooth. Such molars are therefore capable of both shearing and grinding, and provide more versatile occlusal functions than the primitive (for example, “symmetrodont”) teeth that have simple, wedge-like trigonids limited to shearing. The increased versatility of this functional complex is considered to be the precursor condition to many mammalian dental specializations that are correlated with the adaptive radiations of marsupials and placentals³⁴.

Evolution of tribosphenic molars has been a focal point for interpreting early mammal history, owing to influential studies on cusp homologies^{3,35} and the evolution of wear facets⁴, postulating a transformation of the elaborated tribosphenic occlusion through a stepwise assembly of discrete dental characters with seemingly simple wear functions. On the basis of these fundamental under-

standings, the tribosphenic complex with mortar-to-pestle function has been considered to be a ‘hallmark’ apomorphy distinguishing marsupials, placentals and their Laurasian fossil therian relatives^{1,16}.

The trigonid wedge of the tribosphenic molar is primitive: its basic structure consists mainly of plesiomorphies shared by non-tribosphenic holotheres^{1,3,4,16–19,35}. Only the mortar-to-pestle grinding features on the talonid heel comprise the apomorphies that define the tribosphenic condition (where upper molars are not known, the presence of a protocone is inferred on the basis of talonid structure and wear surfaces^{6,7}). The traditional view that such a highly sophisticated functional complex could evolve only once has been contradicted by the discovery of the unique holotherian *Shuotherium*¹⁹, in which the pseudo-talonid basin (mortar) is developed anterior to the trigonid, opposite to the posterior talonid of tribosphenic mammals^{20,21}. The main lingual cusp (“pseudo-protocone”²⁰) of *Shuotherium* upper molars could nonetheless occlude with this mortar of the pseudo-talonid^{20,21}, and is functionally analogous to the tribosphenic condition. This homoplasy provides a plausible case that evolution of the mortar-to-pestle functional complex occurred more than once among holotherian mammals.

Our study corroborates the interpretation that *Ambondro* has a fully functional talonid basin typical of tribosphenic molars⁶. *Steropodon*, as with *Obdurodon*¹⁴, lacks mortar-to-pestle occlusion^{13,14,22} on its talonid. We concur with the observation that *Steropodon* and *Obdurodon* have an elevated talonid floor¹⁴, and thus differ from the earliest known, northern tribosphenic mammals in which the talonid floor is lower and more excavated, despite their convergent resemblance in other aspects of the talonid. Given the evidence for monophyletic grouping of *Steropodon*, *Ambondro* and *Ausktribosphenos*, it is equally plausible that the mortar-to-pestle occlusion on the talonid developed independently in *Ausktribosphenos*^{7,8} and *Ambondro*⁶ in australosphenidans (Fig. 1a), or that the hypertrophied transverse crests evolved secondarily in place of a functional basin on the talonid in *Steropodon*¹³ and the earliest ornithorhynchid *Obdurodon*^{13,14,28}. Correspondingly, the hypertrophied crests also developed secondarily in place of a functional protocone on the upper molars, as seen in *Obdurodon*^{13,14}. These autapomorphic molar features of *Steropodon* and ornithorhynchids represent a convergent pattern that has occurred in many unrelated lineages in marsupials, such as diprotodontoids, and in placentals, such as perissodactyls.

The classical theory on functional evolution of tribosphenic molars^{3,4} and the monophyly hypothesis of tribosphenic mammals¹ were developed on the basis of the available fossil record up to the early 1970s, when tribosphenic mammals were known exclusively from the Laurasian continents, and long before their relatively recent discoveries in Gondwana. The precociously derived talonid features^{13–15} and great antiquity⁶ of australosphenidan mammals are most parsimoniously explained, on the basis of present evidence, if the functional evolution of the mortar-to-pestle molar occlusion, as originally formulated⁴, occurred more than once. Our analysis not only supports the long-established clade of the northern tribosphenic mammals (“tribosphenidans”^{1,12,16}), but also suggests that the newly discovered southern tribosphenic mammals form a monophyletic clade in their own right.

Palaeobiogeography

The earliest tribosphenic mammals (earliest Cretaceous) were originally known only from Eurasia, or an adjacent—and biogeographically related—part of northwestern Africa^{36,37}. The earliest eutherians are from the late Early Cretaceous of Asia¹⁰ and, perhaps, North America¹¹; similarly, early metatherians are known from these landmasses^{12,38}. These distributions have been viewed as consistent with the hypothesis that tribosphenidans and subgroups thereof⁵ have a northern origin.

Recent discoveries have led to a different interpretation.

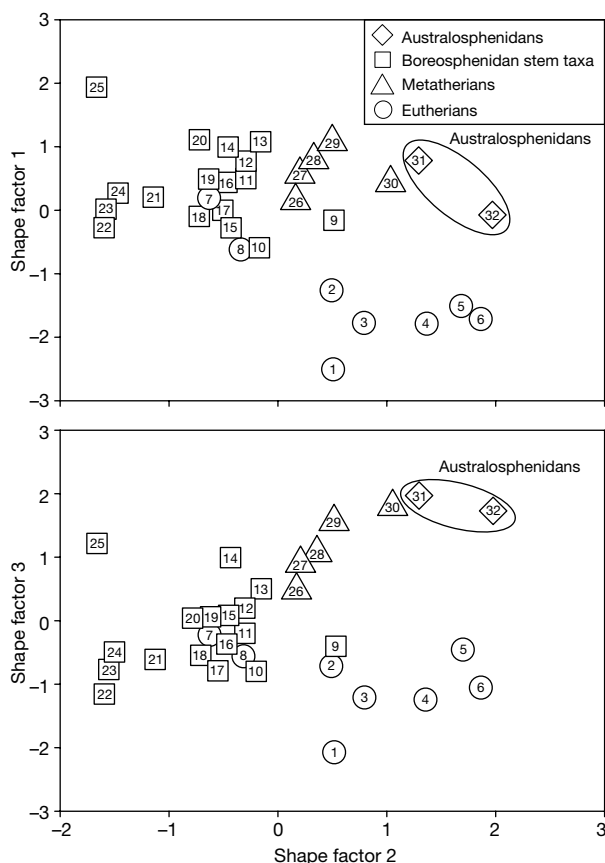


Figure 3 Lower molar shape in early mammals. Morphometric analysis (see Methods) shows australosphenidan taxa to be distinct from other Mesozoic tribosphenic mammals in the shape of their lower molars. Circles, stem taxa of boreosphenidans (eupantothere and ‘tribotheres’): 1, *Peramus*; 2, *Kermackia*; 3, *Hypomylos*; 4, *Kielantherium*; 5, *Potamotelses*; 6, *Deltatheridium*; 7, *Holoclemensia*; 8, *Pappotherium*. Squares, eutherians: 9, *Prokennalestes* spp.; 10, *Montanalestes*; 11, *Gypsonictops hypoconus*; 12, *Paranyctoides* spp.; 13, *Purgatorius unio*; 14, *Gallolestes pachymandibularis*; 15, *Procerberus*; 16, *Otlestes*; 17, *Batodon*; 18, *Kennalestes*; 19, *Asioryctes*; 20, *Protungulatum* spp.; 21, *Cimolestes cerberoides*; 22, *Cimolestes propalaeoryctes*; 23, *Cimolestes incisus*; 24, *Cimolestes magnus*; 25, *Zalambdalestes*. Triangles, metatherians: 26, *Kokopellia*; 27, *Pediomys elegans*; 28, *Glasbius intricatus*; 29, *Didelphodon vorax*; 30, *Alphadon marshi*. Diamonds, southern australosphenidan taxa: 31, *Ausktribosphenos*; 32, *Steropodon*. *Ambondro* is omitted because of incompleteness.

Ambondro, from the Middle Jurassic of Madagascar, is at least 25 Myr older than any previously reported tribosphenic mammals⁶; *Ausktribosphenos*, from the late Early Cretaceous of Australia^{7,8}, is nearly as old as the earliest northern therians. This early presence of tribosphenic mammals in Gondwana has led to the proposal that they originated on the southern continents^{6–8,37}. Each of the preceding biogeographical hypotheses assumed that the apomorphic, mortar-to-pestle structure of tribosphenic molars had a monophyletic origin.

We propose a new model: mammals with the mortar-to-pestle structure of the talonid (tribosphenic condition) underwent vicariant and diphyletic evolution on Gondwanan and Laurasian continents, respectively (Fig. 1). We suggest that the Australosphenida represent an endemic radiation of mammals on southern continents during the Jurassic/Early Cretaceous, surviving through the Tertiary^{22,28,39} as the extant monotremes.

A diphyletic origin and early vicariant evolution of mammals with tribosphenic dentitions is consistent with the well-documented provincialism of mammals in the Mesozoic of South America, which (before the latest Cretaceous) includes highly distinctive, endemic taxa related to Jurassic forms of northern continents^{39–41}. The known record suggests that boreosphenidans and their subgroups were largely or wholly confined to northern continents during much of their Mesozoic history, until a later dispersal of marsupials and placentals into South America^{6,42} and other southern continents occurred during the Late Cretaceous/Early Tertiary.

Systematic palaeontology

Class Mammalia Linnaeus 1758

Subclass Holotheria Wible *et al.* 1995 (ref. 24)

Infraclass Australosphenida nov.

(Include Monotremata; Ausktribosphenida^{7,8}; and *Ambondro*⁶).

Etymology. From *australis* (Latin, meaning southern) and *sphen* (Greek, meaning wedge), referring to their origin on the Gondwanan continents during the Jurassic and to the tribosphenic molars.

Diagnosis. Holotherians^{2,24} with tribosphenic molars; differ from boreosphenidans in having a continuous, shelf-like mesial cingulid that extends to the lingual side of the lower molar, rather than individualized cingulid cusps, and ultimate lower premolar with fully triangulated trigonid, a derived condition unseen in any other holotherians of the Jurassic/Early Cretaceous; differ from all non-tribosphenic holotherians in having a transversely wide talonid; differ from *Shuotherium* in having talonid placed posterior to the trigonid; and differ from boreosphenidans (including extant therians) by primitive retention, at least in *Ausktribosphenos*, of the postdentary trough on the dentary (also see Supplementary Information).

Distribution. Middle Jurassic of Madagascar; Early Cretaceous to recent of Australian region; Palaeocene of South America. The oldest australosphenidan is *Ambondro* from the Middle Jurassic of Madagascar.

Infraclass Boreosphenida, nov.

(include Tribosphenida McKenna, 1975 (ref. 1))

Etymology. From *boreas* (Latin, meaning northern wind) and *sphen* (Greek, meaning wedge), referring to their origin on the northern continents during the Mesozoic and to tribosphenic molars; meaning Tribosphenida¹ of the northern continents.

Revised diagnosis. Differ from all non-boreosphenidans in the posterior placement of the mandibular angle (to the level of dentary condyle); differ from kuehneotheriid holotherians^{16,43}, some eupantotheres and australosphenidans by absence of the primitive postdentary trough for the postdentary bones (except coronoid) in the dentary; differ from all mammals except Australosphenida by presence of tribosphenic molars; differ from *Shuotherium* in having talonid placed posterior to the trigonid in the lower molars, but convergent to the latter's pseudo-protcone of the upper molar; differ from Australosphenida by having distinctive

cingulid cusps but lacking the continuous mesial cingulid on the molars of the latter, and lacking the triangulated trigonid on the ultimate lower premolar in the Early Cretaceous taxa (also see Supplementary Information).

Distribution. Restricted to the Northern Hemisphere during the Early Cretaceous; present in latest Cretaceous of South America, India and northern continents; Tertiary to recent of the world. The oldest boreosphenidans are Berriasian in age^{36,37}.

Methods

Hypothesis of relationships among tribosphenic mammals is based on PAUP (3.1.1, Swofford, 1993) analyses of 21 selected taxa (see Supplementary Information). Our primary analysis (Fig. 1a) concentrated on 55 dental and mandibular characters that are preserved in many southern taxa (*Ambondro*, *Ausktribosphenos* and *Steropodon*). We grouped the southern and northern tribosphenic mammals into independent clades, by the strict consensus of 18 equally parsimonious and shortest trees. Our extended analysis (Fig. 1b) added 70 cranial and postcranial characters (from refs 23–27) that are preserved in the more complete taxa in our dataset to the 55 dental and mandibular characters known for the southern mammals, with requisite missing values coded for the taxa known only by teeth and jaws. The australosphenidan clade (including monotremes) and boreosphenidan clade (including extant therians) are supported by simultaneous analyses of 118 parsimony-informative characters of dentition, skull and skeleton for the well-preserved taxa of early mammals^{23–27} (Fig. 1b), and with 125 characters in additional tests of 26 taxa.

Diagnoses and relative strength of tree nodes

The australosphenidan clade is a more robust node (that takes more steps to collapse) than many widely accepted clades, such as the boreosphenidan clade (*Kielantherium* plus extant therians), the extant therian clade and the eutherian clade (Fig. 1). Apomorphies for each node of the cladograms are listed in Supplementary Information.

Additional tests

Both the australosphenidan and the boreosphenidan clades are maintained if five more taxa (tritilodontids, *Obdurodon*, *Ornithorhynchus*, multituberculates and *Vincelestes*) are added to the simultaneous analysis of 125 dental, mandibular, cranial and postcranial datasets (see Supplementary Information). The extant platypus (*Ornithorhynchus*) is nested in the australosphenidan clade. Bootstrap majority consensus trees from separate searches of the 55 dental and mandibular data partitions and the total 125 morphological characters have replicated both the australosphenidan clade and the boreosphenidan clade (Supplementary Information). These analyses (Fig. 1) and other independent morphological analyses^{2,23–27} have all shown that monotremes (herein considered to be a part of the australosphenidan clade) are more distant from the living therians than the non-tribosphenic symmetrodonts (*Zhangheotherium*) and eupantotheres (*Henkelotherium*).

Multivariate shape analysis

Plots in Fig. 3 show the first three principal components of principal components analysis, in which data have been standardized and variation owing to size has been removed. Positive scores for the first component reflect an anteroposteriorly compressed, narrow trigonid; for the second component, wide talonid and longer distance from hypoconid to hypoconulid; and for the third, a short metaconid and (of lesser importance) protoconid. Factor 1 distinguishes the southern australosphenidan taxa (*Ausktribosphenos*, *Steropodon*) from stem boreosphenidans, whereas factors 2 and 3 distinguish them from marsupials and eutherians. Methods for data collection and multivariate analysis are given in refs 11 and 44.

Received 15 February; accepted 6 November 2000.

- McKenna, M. C. in *Phylogeny of the Primates* (eds Luckett, W. P. & Szalay, F. S.) 21–46 (Plenum, New York, 1975).
- Hopson, J. A. in *Major Features of Vertebrate Evolution* (ed. Spencer, R. S.) 190–219 (The Paleontological Society, Knoxville, 1994).
- Patterson, B. Early Cretaceous mammals and the evolution of mammalian molar teeth. *Fieldiana* **13**, 1–105 (1956).
- Crompton, A. W. in *Early Mammals* (eds Kermack, D. M. & Kermack, K. A.) 65–87 (Zool. J. Linn. Soc., London, 1971).
- Lillegraven, J. A. Biogeographical considerations of the marsupial-placental dichotomy. *Annu. Rev. Ecol. Syst.* **5**, 263–283 (1974).
- Flynn, J. J., Parrish, J. M., Rakotosamimanana, B., Simpson, W. F. & Wyss, A. E. A Middle Jurassic mammal from Madagascar. *Nature* **401**, 57–60 (1999).
- Rich, T. H. *et al.* A tribosphenic mammal from the Mesozoic of Australia. *Science* **278**, 1438–1442 (1997).
- Rich, T. H. *et al.* Early Cretaceous mammals from Flat Rocks, Victoria, Australia. *Rec. Queen Vict. Mus.* **106**, 1–35 (1999).
- Kielan-Jaworowska, Z., Cifelli, R. L. & Luo, Z. Alleged Cretaceous placental from down under. *Lethaia* **31**, 267–268 (1998).
- Kielan-Jaworowska, Z. & Dashzeveg, D. D. Eutherian mammals from the Early Cretaceous of Mongolia. *Zool. Scripta* **18**, 347–355 (1989).
- Cifelli, R. L. Tribosphenic mammal from the North American Early Cretaceous. *Nature* **401**, 363–366 (1999).
- Rougier, G. W., Wible, J. R. & Novacek, M. J. Implications of *Deltatheridium* specimens for early marsupial history. *Nature* **396**, 459–463 (1998).

13. Archer, M., Flannery, T. F., Ritchie, A. & Molnar, R. First Mesozoic mammal from Australia—an Early Cretaceous monotreme. *Nature* **318**, 363–366 (1985).
14. Archer, M., Murray, P., Hand, S. J. & Godthelp, H. in *Mammal Phylogeny* (eds Szalay, F. S., Novacek, M. J. & McKenna, M. C.) (Springer, New York, 1993).
15. Musser, A. M. & Archer, M. New information about the skull and dentary of the Miocene platypus *Obdurodon dicksoni*, and a discussion of ornithorhynchid relationships. *Phil. Trans. R. Soc. Lond. B* **353**, 1063–1079 (1998).
16. Prothero, D. R. New Jurassic mammals from Como Bluff, Wyoming, and the interrelationships of non-tribosphenic Theria. *Bull. Am. Mus. Nat. Hist.* **167**, 277–326 (1981).
17. Clemens, W. A. Jr & Mills, J. R. E. Review of *Peramus tenuirostris*. *Bull. Br. Mus. Nat. Hist.* **20**, 89–113 (1971).
18. Krebs, B. Das Skelett von *Henkelotherium guimarotae* gen. et sp. nov. (Eupantotheria, Mammalia) aus dem Oberen Jura von Portugal. *Berliner Geowiss. Abh.* **133**, 1–121 (1991).
19. Chow, M. & Rich, T. H. *Shuotherium dongi*, n. gen. and sp., a therian with pseudo-tribosphenic molars from the Jurassic of Sichuan, China. *Aust. Mammal.* **5**, 127–142 (1982).
20. Wang, Y.-Q., Clemens, W. A., Hu, Y.-M. & Li, C.-K. A probable pseudo-tribosphenic upper molar from the Late Jurassic of China and the early radiation of the Holotheria. *J. Vert. Paleontol.* **18**, 777–787 (1998).
21. Sigogneau-Russell, D. Discovery of a Late Jurassic Chinese mammal in the upper Bathonian of England. *C. R. Acad. Sci.* **327**, 571–576 (1998).
22. Kielan-Jaworowska, Z., Crompton, A. W. & Jenkins, F. A. Jr The origin of egg-laying mammals. *Nature* **326**, 871–873 (1987).
23. Rougier, G. W., Wible, J. R. & Hopson, J. A. Basicranial anatomy of *Priacodon fruitaensis* (Triconodontidae, Mammalia) from the Late Jurassic of Colorado, and a reappraisal of mammaliaform interrelationships. *Am. Mus. Novit.* **3183**, 1–38 (1996).
24. Wible, J. R., Rougier, G. W., Novacek, M. J., McKenna, M. C. & Dashzeveg, D. D. A mammalian petrosal from the Early Cretaceous of Mongolia: implications for the evolution of the ear region and mammalian relationships. *Am. Mus. Novit.* **3149**, 1–19 (1995).
25. Rowe, T. B. Definition, diagnosis, and origin of Mammalia. *J. Vert. Paleontol.* **8**, 241–264 (1988).
26. Hu, Y.-M., Wang, Y.-Q., Luo, Z.-X. & Li, C.-K. A new symmetrodont mammal from China and its implications for mammalian evolution. *Nature* **390**, 137–142 (1997).
27. Ji, Q., Luo, Z.-X. & Ji, S.-A. A Chinese triconodont mammal and mosaic evolution of the mammalian skeleton. *Nature* **398**, 326–330 (1999).
28. Archer, M. *et al.* The evolutionary history and diversity of Australian mammals. *Aust. Mammal.* **21**, 1–45 (1999).
29. Griffiths, M. *The Biology of the Monotremes* (Academic, New York, 1978).
30. Zeller, U. Die Entwicklung und Morphologie des Schädels von *Ornithorhynchus anatinus* (Mammalia: Prototheria: Monotremata). *Abh. der Senckenb. Naturfor. Gesel.* **545**, 1–188 (1989).
31. Killian, J. K. *et al.* M6P/IGF2R imprinting evolution in mammals. *Mol. Cell* **5**, 707–716 (2000).
32. Penny, D. & Hasegawa, M. The platypus put in its place. *Nature* **387**, 549–550 (1997).
- Janke, A., Xu, X., & Anason, U. The complete mitochondrial genome of wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proc. Natl Acad. Sci. USA* **94**, 1276–1281 (1997).
34. Jacobs, L. L., Winkler, D. A. & Murry, P. A. Modern mammal origins: evolutionary grades in the Early Cretaceous of North America. *Proc. Natl Acad. Sci. USA* **86**, 4992–4995 (1989).
35. Butler, P. M. The teeth of the Jurassic mammals. *Proc. Zool. Soc. Lond.* **109**, 329–356 (1939).
36. Sigogneau-Russell, D. & Ensom, P. C. Découverte, dans le groupe de Purbeck (Berriasien, Angleterre), de plus ancien témoignage de l'existence de mammifères tribosphéniques. *C. R. Acad. Sci.* **319**, 833–838 (1994).
37. Sigogneau-Russell, D. Découverte du premier mammifère tribosphénique du Mésozoïque africain. *C. R. Acad. Sci.* **313**, 1635–1640 (1991).
38. Cifelli, R. L. & de Muizon, C. Dentition and jaw of *Kokopellia juddi*, a primitive marsupial or near marsupial from the medial Cretaceous of Utah. *J. Mammal. Evol.* **4**, 241–258 (1997).
39. Pascual, R. *et al.* First discovery of monotremes in South America. *Nature* **356**, 704–705 (1992).
40. Bonaparte, J. F. New Late Cretaceous mammals from the Los Alamos Formation, northern Patagonia. *Nat. Geogr. Res.* **6**, 63–93 (1990).
41. Pascual, R., Goin, F. J., González, P., Ardolino, A. & Puerta, P. A highly derived docodont from the Patagonian Late Cretaceous: evolutionary implications for Gondwanan mammals. *Geodiversitas* **22**, 395–414 (2000).
42. de Muizon, C., Cifelli, R. L. & Céspeles, R. The origin of dog-like marsupials and the early evolution of Gondwanian marsupials. *Nature* **389**, 486–489 (1997).
43. McKenna, M. C. & Bell, S. K. *Classification of Mammals Above the Species Level* (Columbia Univ. Press, New York, 1997).
44. Butler, P. M. Early trends in the evolution of tribosphenic molars. *Biol. Rev.* **65**, 529–552 (1990).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank T. H. Rich, P. Vickers-Rich and M. Archer for providing comparative casts; and K. C. Beard, A. W. Crompton, M. R. Dawson, J. J. Flynn, J. J. Hurum, M. C. McKenna, M. J. Novacek, R. Presley, T. H. Rich, D. Sigogneau-Russell, J. R. Wible and A. R. Wyss for relevant discussion, and M. Klingler for assistance in illustration. Research was supported by National Science Foundation (USA) and National Geographic Society (Z.-X. L. and R.L.C.), Carnegie Museum (Z.-X.L.) and Institute of Paleobiology, PAN (Z.K.-J.).

Correspondence and requests for materials should be addressed to R.L.C. (e-mail: rlc@ou.edu).

3-
3.

A massive reservoir of low-excitation molecular gas at high redshift

Padeli Papadopoulos*, Rob Ivison†, Chris Carilli‡ & Geraint Lewis||

* Leiden Observatory, PO Box 9513, 2300 Leiden, The Netherlands

† Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

‡ National Radio Astronomy Observatories, PO Box 0, Socorro, New Mexico 87801-0387, USA

|| Anglo-Australian Observatory, PO Box 296, Epping, NSW 1710, New South Wales, Australia

Molecular hydrogen (H_2) is an important component of galaxies because it fuels star formation and the accretion of gas onto active galactic nuclei (AGN), the two processes that can generate the large infrared luminosities of gas-rich galaxies^{1,2}. Observations of spectral-line emission from the tracer molecule carbon monoxide (CO) are used to probe the properties of this gas. But the lines that have been studied in the local Universe—mostly the lower rotational transitions of $J = 1 \rightarrow 0$ and $J = 2 \rightarrow 1$ —have hitherto been unobservable in high-redshift galaxies. Instead, higher transitions have been used, although the densities and temperatures required to excite these higher transitions may not be reached by much of the gas. As a result, past observations may have underestimated the total amount of molecular gas by a substantial amount. Here we report the discovery of large amounts of low-excitation molecular gas around the infrared-luminous quasar APM08279+5255 at redshift $z = 3.91$, using the two lowest excitation lines of ^{12}CO ($J = 1 \rightarrow 0$ and $J = 2 \rightarrow 1$). The maps confirm the presence of hot and dense gas near the nucleus³, and reveal an extended reservoir of molecular gas with low excitation that is 10 to 100 times more massive than the gas traced by the higher-excitation observations. This raises the possibility that significant amounts of low-excitation molecular gas may exist in the environments of high-redshift ($z > 3$) galaxies.

APM08279+5255 was discovered serendipitously during a search for Galactic halo carbon stars⁴. It comprises a strongly lensed broad absorption line quasar at redshift $z = 3.9$ with a complex optical spectrum and a bolometric luminosity of $\sim 10^{14} L_\odot$ (where $L_\odot$ is the luminosity of the Sun), after correcting for lensing magnification; it displays extremely strong rest-frame far-infrared emission⁵ suggestive of dust-rich gas undergoing rapid star formation. The CO($4 \rightarrow 3$) and CO($9 \rightarrow 8$) lines have been detected³ suggesting that the molecular gas must be warm, extended on a scale of up to $1''$ in the image plane (only 160–270 pc in the source plane), with a mass of $(1\text{--}6) \times 10^9 M_\odot$ (where $M_\odot$ is the mass of the Sun).

We have used the National Radio Astronomy Observatory's Very Large Array (VLA) to detect and image CO $J = 1 \rightarrow 0$ and $J = 2 \rightarrow 1$ emission from APM08279+5255. Emission in the CO($1 \rightarrow 0$) line (Fig. 1) extends over scales of $\sim 7''$ (~ 30 kpc at $z = 3.9$ for $H_0 = 75 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $q_0 = 0.5$), well beyond the gravitationally lensed nucleus⁶ where the emission from the CO($4 \rightarrow 3$) and CO($9 \rightarrow 8$) transitions is confined³. Maps of CO($2 \rightarrow 1$) detect the northern part of the extended emission and show it to consist of two kinematically and spatially distinct features (Fig. 2). The continuum emission detected at 23 GHz emanates from the lensed nucleus and is unresolved ($\leq 0.6''$) with a flux density of $0.30 \pm 0.05 \text{ mJy}$. At 43 GHz, no continuum emission is detected at a level of $\sigma = 0.30 \text{ mJy}$, and the emission expected from the hot dust in the nucleus⁵ is negligible ($\leq 0.04 \text{ mJy}$). No appreciable dust continuum is expected underlying the extranuclear CO($2 \rightarrow 1$) emission from the cooler dust in those regions.

The nuclear CO($1 \rightarrow 0$) line emission has a velocity-integrated flux

density of $0.150 \pm 0.045 \text{ Jy km s}^{-1}$ and a deconvolved size of $\sim 0.6''\text{--}1''$, confirmed by the CO($2 \rightarrow 1$) map (Fig. 2) which shows this emission extending $\sim 0.8''$ along the direction expected from matched-resolution 1.6- μm Hubble Space Telescope (HST) images⁶. This size is similar to that of the CO($4 \rightarrow 3$) and CO($9 \rightarrow 8$) emission³, while the velocity-averaged brightness temperature ratios, $R_{4 \rightarrow 3} = (4 \rightarrow 3)/(1 \rightarrow 0)$ and $R_{2 \rightarrow 1} = (2 \rightarrow 1)/(1 \rightarrow 0)$, are 1.5 ± 0.5 and 1.35 ± 0.55 , respectively. These values are perfectly compatible with a hot and dense gas phase in the nucleus.

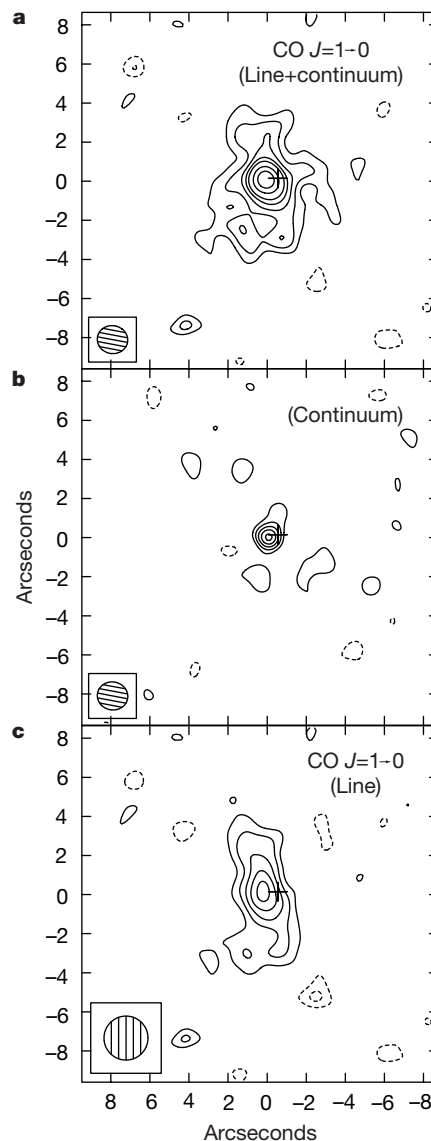


Figure 1 Molecular gas in and around APM08279+5255, as traced by CO $J = 1 \rightarrow 0$, and continuum emission from the active nucleus. **a**, Naturally weighted maps at $\nu_{\text{obs}} = 23.4649 \text{ GHz}$; **b**, $\nu_{\text{obs}} = 23.3649 \text{ GHz}$, corresponding to CO $J = 1 \rightarrow 0$ ($\nu_{\text{rest}} = 115.2712 \text{ GHz}$) and its adjacent continuum, respectively, at $z = 3.912$. The maps were produced from the Fourier transform of the measured visibilities, tapered to enhance low-brightness extended emission. The dirty beam, identical in both maps with full-width at half-maximum (FWHM) $1.5'' \times 1.4''$, is shown bottom left. Contours are at $-3, -2, 2, 3, 4, 5, 6, 8, 10 \times \sigma$, where $\sigma = 40 \mu\text{Jy}$ per beam. **c**, Naturally weighted map of CO $J = 1 \rightarrow 0$ with the continuum source subtracted. The subtraction was performed in the visibility domain and the resulting uv data were then Fourier transformed, CLEANed and convolved to a circular beam, FWHM $2.25''$, shown bottom left. Contours this time are at $-3, -2, 2, 3, 4, 5, 6 \times \sigma$. The cross marks the position of lens component 'A' at RA 08 h 31 min 41.64 s and dec. $+52^\circ 45' 17.5''$ (J2000). The $0.5''$ offset between the peak radio intensity and 'A' is probably caused by HST astrometric errors.

We have considered in detail, and rejected, gravitational lensing as the cause of the extended CO(2 → 1) structure. There are no obvious optical counterparts and the CO(2 → 1) maps demonstrate that the two features to the north and northeast and the quasar nucleus do not share a common CO(2 → 1) line profile, a necessary characteristic if a single galaxy underlies them all. Detailed modelling of the lens^{6,7}, based on high-resolution HST and Keck data, does not suggest the existence of additional lensed features. We thus conclude that the extranuclear features are due to companion galaxies near APM08279+5255. Moreover, their large angular distance from any central lens caustics precludes significant magnification (when an elliptical mass distribution is used to represent the lensing galaxy⁶, which accurately recovers the quasar image configuration, the magnification in this region is ≤ 1.2) and hence the implied molecular gas masses are large.

X_{CO} is the CO-to-H₂ conversion factor^{8–10}, ranging from $\sim 1 M_{\odot} (\text{K km s}^{-1} \text{pc}^2)^{-1}$ in ultraluminous infrared galaxies¹¹ (ULIRGs) to $\sim 5 M_{\odot} (\text{K km s}^{-1} \text{pc}^2)^{-1}$ for quiescent spirals. The integrated flux density of the two extranuclear features in the CO(2 → 1) maps is $1.15 \pm 0.54 \text{ Jy km s}^{-1}$ which (for the aforementioned range of X_{CO} and $R_{2-1} \approx 1$) yields a mass of $M(\text{H}_2) \approx (0.65\text{--}3.2) \times 10^{11} M_{\odot}$, this value being merely a lower limit if sub-thermal excitation of

CO(2 → 1) ($R_{2-1} < 1$) or sub-solar metallicities ($X_{\text{CO}} > 5 M_{\odot} (\text{K km s}^{-1} \text{pc}^2)^{-1}$) are important.

This molecular gas reservoir is $\sim 10\text{--}100$ times more massive than that deduced for the nuclear region³ and neither CO(4 → 3), CO(9 → 8) nor high-resolution 210-GHz continuum maps give any hint of it. This mass is comparable to that of an entire spiral galaxy and is confined within $2''\text{--}3''$ or $\sim 10 \text{ kpc}$ (Fig. 2). Similar gas-rich companions with no obvious optical counterparts, akin to the recently discovered submillimetre galaxy population¹², have been found in the vicinity of other high-redshift systems^{13,14}. They have the tell-tale marks of intense (far-infrared luminosity $L_{\text{FIR}} \approx 10^{13} L_{\odot}$), heavily obscured starbursts in merger configurations, converting a large reservoir of molecular gas ($\sim 10^{11} M_{\odot}$) into stars. In these systems the dynamical mass is comparable to the gas mass, a situation possible for the companion galaxies to APM08279+5255, though only observations with higher spatial resolution and wider velocity coverage can address this issue properly.

The failure of sensitive spectral-line observations to detect any extranuclear CO(4 → 3) and CO(9 → 8) emission implies low values (≤ 0.15) for the global R_{4-3} and R_{9-8} ratios for the massive reservoirs of molecular gas near APM08279+5255. This has far-reaching consequences: at $z > 3$, millimetre-wave telescopes can only access CO $J + 1 \rightarrow J, J > 2$ and hence cannot hope to trace the bulk of the H₂ gas.

The low excitation of lines such as CO(9 → 8) is hardly surprising, even for a starburst environment, because the average conditions in the molecular gas will not be sufficient for their large-scale excitation. If the physical conditions in Orion A¹⁵ are typical of a starburst environment, $R_{9-8} \leq 0.15$ (and in most cases $R_{9-8} \leq 0.05$). Moreover, the molecular interstellar medium in intense starburst environments may not be just an ensemble of warm and dense Orion-type clouds. Where the most energetic star formation takes place, the gas may also have a warm but diffuse phase^{11,16}, $n(\text{H}_2) \leq 10^3 \text{ cm}^{-3}$, generated by the disruptive effects of intense ultraviolet radiation, supernovae and the violent dynamics associated with mergers. The CO $J + 1 \rightarrow J, J > 2$ emission from such gas will be weak, possibly for the entire starburst if this phase is dominant. Starbursts with similar far-infrared and low- J CO luminosities can be dominated by drastically different gas phases^{17,18}. Such differences are expected to occur during the evolution of a starburst, depending on the complex interplay between the accumulation of dense gas that feeds the star formation (for example, through a merger) and the cloud disruption caused by the products of that process, the massive stars.

Finally, since CO observations at high redshift usually sample emission averaged over several kiloparsecs, the possibility of cold and diffuse gas dominating the global CO emission should not be discounted. The CO $J + 1 \rightarrow J, J > 2$ transitions may then have low excitation and thus be very faint, a state not altered by the enhanced cosmic microwave background¹⁴. The detection of CO in galaxies at $z > 3$ means that significant metal-enrichment has taken place; low-excitation gas would then represent a previously metal-enriched part of their interstellar medium that is not participating in the observed star-formation episode. Such excitation gradients have been found in several nearby infrared-luminous galaxies hosting powerful circumnuclear starbursts and AGN¹⁹.

APM08279+5255 can be seen as an extreme case of such an excitation gradient, further exaggerated by gravitational lensing of a particularly hot region of the underlying system. Differential lensing seems to be preferentially boosting the far-infrared and CO emission of the AGN-heated (rather than the starburst-heated) component of the interstellar medium. This is strongly suggested by the high temperature of the gas and dust ($\sim 200 \text{ K}$) which is not typical of nearby starbursts²⁰ or of clouds heated by young stars (for example, Orion A, $T \approx 40\text{--}60 \text{ K}$). Also, the ratio $L_{\text{FIR}}/M(\text{H}_2) \approx 10^4 L_{\odot} M_{\odot}^{-1}$ is about 40 times higher than the highest values observed in ULIRGs, and the $L_{\text{FIR}}/L_{1.4 \text{ GHz}}$ ratio is an order of

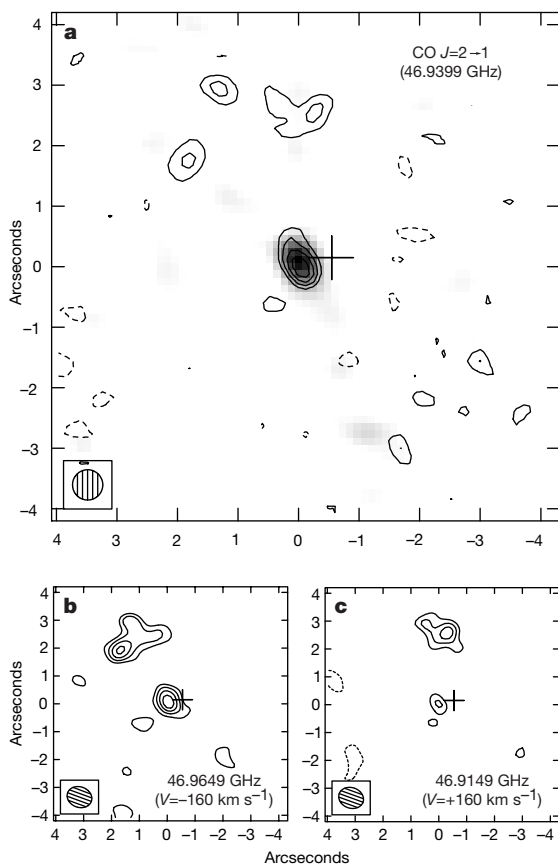


Figure 2 Molecular gas in and around APM08279+5255, as traced by CO $J = 2 \rightarrow 1$, overlaid on a greyscale image of 8.45-GHz continuum emission from the active nucleus. **a**, Naturally weighted map of CO $J = 2 \rightarrow 1$ ($\nu_{\text{rest}} = 230.5380 \text{ GHz}$) with both IF bands averaged (contours) and overlaid on continuum emission at 8.45 GHz (greyscale) at a common resolution ($0.5'' \times 0.5''$). Contours are at $-3, -2, 2, 3, 4, 5 \times \sigma$, where $\sigma = 0.15 \text{ mJy per beam}$ and the greyscale range is $25(2\sigma)\text{--}305 \mu\text{Jy per beam}$. CLEAN was applied. **b, c**, Naturally weighted, tapered maps of CO $J = 2 \rightarrow 1$ for the two IF bands separately, at a common resolution of $0.90'' \times 0.75''$ FWHM. Contours are at $-3, 3, 4, 5, 6 \times \sigma$, where $\sigma = 0.20 \text{ mJy per beam}$. CLEAN was not applied. In all cases, the beam FWHM is shown bottom left and the cross marks the position of lens⁶ component 'A'.

magnitude larger than in star-forming galaxies. This all suggests a heating source other than stars and the AGN is the obvious alternative. Differential magnification of AGN-heated gas therefore seems to be responsible for the high luminosity of the CO(9 → 8) line.

Although highly excited regions with CO emission amplified by gravitational lensing are good markers of the presence of molecular gas at high redshifts, they may give a poor representation of its average physical conditions, particularly of its total mass and distribution. At $z > 3$, this excitation bias becomes more serious since millimetre-wave interferometers, the instruments usually used for such searches, can detect only CO $J + 1 \rightarrow J, J > 2$ whose excitation requirements are $n(\text{H}_2) \geq 10^4 \text{ cm}^{-3}$ and $T \geq 50 \text{ K}$. These values, being typical of the average conditions in star-forming clouds, mark a gradual excitation turnover.

Observations of the CO(1 → 0) and CO(2 → 1) transitions, with their minimal excitation requirements, may reveal much larger molecular gas reservoirs at high redshifts. The VLA is currently the only instrument capable of sensitive, sub-arcsecond observations of these lines and will thus be an important tool for unbiased surveys of metal-enriched H_2 gas around objects in the distant Universe²¹. □

Methods

Our observations took place during 23–24 April 2000 using the VLA in its C configuration. After all overheads, 20 hours were spent integrating on APM08279+5255. 19 of the 27 antennas were equipped with 43-GHz receivers ($T_{\text{sys}} = 45\text{--}100 \text{ K}$); all were equipped with 22-GHz receivers, including eleven new receivers with $T_{\text{sys}} \sim 40 \text{ K}$ (compared with 100–150 K for the remainder).

We used the new fast-switching technique^{22,23}, recording data every 3.3 s, with 30 s on the calibrator and 170 s on the source (210 s in the 22-GHz band); the typical slewing overhead to the compact phase calibrator, 0824+558, 3.3° away was 7 s. The pointing accuracy was checked every hour.

This technique yields diffraction-limited images at the highest VLA operating frequencies over long baselines. Conventional phase referencing would not have been able to track tropospheric phase variations and self-calibration techniques could not be employed because of the low signal-to-noise ratio per baseline.

We used the continuum mode, placing emphasis on sensitivity rather than line profile information. A dual-polarization 50-MHz band was placed as close to the expected line centre as could be allowed by correlator limitations: 23.4649 GHz for CO $J = 1 \rightarrow 0$ (+91 km s^{-1} from the line centre³). Bandpass roll-off limited the effective bandwidth per intermediate frequency (IF) to ~45 MHz which, at $z = 3.9$, corresponds to $\Delta v \sim 575 \text{ km s}^{-1}$ at 23 GHz. The remaining IF pair was tuned to simultaneously observe the continuum at 23.3649 GHz (+1,280 km s^{-1}).

We obtained matching velocity coverage at 43 GHz (CO $J = 2 \rightarrow 1$) by placing two contiguous 50-MHz dual-polarization bands at 46.9399 GHz (−20 km s^{-1} from the line centre). The continuum was observed on a separate occasion with both IF pairs tuned to 43.3 GHz. The flux-density scale was fixed using 3C286; the uncertainty is ~15% at 23 GHz and ~20% at 46 GHz. Calibration and reduction of the data was standard in most respects and the r.m.s. noise in all the maps is similar to the expected theoretical limit.

We cannot completely rule out the possibility that phase errors—caused by atmospheric fluctuations too fast to be tracked even by our fast-switching scheme—are the cause of the extended CO emission, but we consider it to be remote. This is borne out by an extensive series of tests: inspection of the raw visibilities; separate imaging of left- and right-hand polarization maps; imaging of the most phase-stable subset of the data; and imaging of the calibrator source. The agreement between the nuclear CO $J = 2 \rightarrow 1$ emission and its associated 3.5-cm continuum is also strong testimony to the coherence of the CO $J = 2 \rightarrow 1$ map; the point-like 23.4-GHz continuum emission, observed simultaneously with the neighbouring line emission, provides similar supporting evidence for the resolved CO $J = 1 \rightarrow 0$ emission.

Received 5 September; accepted 13 November 2000.

1. Telesco, C. M. Enhanced star formation and infrared emission in the centers of galaxies. *Annu. Rev. Astron. Astrophys.* **26**, 343–376 (1988).
2. Sanders, D. B., Scoville, N. Z. & Soifer, B. T. Molecular gas in luminous infrared galaxies. *Astrophys. J.* **370**, 158–171 (1991).
3. Downes, D., Neri, R., Wiklund, T., Wilner, D. J. & Shaver, P. A. Detection of CO(4–3), CO(9–8), and dust emission in the broad absorption line quasar APM 08279+5255 at a redshift of 3.9. *Astrophys. J.* **513**, L1–L4 (1999).
4. Irwin, M. J., Ibat, R. A., Lewis, G. F. & Totten, E. J. APM 08279+5255: an ultraluminous broad absorption line quasar at a redshift $z = 3.87$. *Astrophys. J.* **505**, 529–535 (1998).
5. Lewis, G. F., Chapman, S. C., Ibat, R. A., Irwin, M. J. & Totten, E. J. Submillimeter observations of the ultraluminous broad absorption line quasar APM 08279+5255. *Astrophys. J.* **505**, L1–L5 (1998).
6. Ibat, R. A., Lewis, G. F., Irwin, M. J., Lehár, J. & Totten, E. J. NICMOS and VLA observations of the gravitationally lensed ultraluminous BAL quasar APM 08279+5255: detection of a third image. *Astron. J.* **118**, 1922–1930 (1999).

7. Egami, E. *et al.* APM 08279+5255: Keck near- and mid-infrared high-resolution imaging. *Astrophys. J.* **535**, 561–574 (2000).
8. Young, J. S. & Scoville, N. Z. Extragalactic CO—gas distributions which follow the light in IC 342 and NGC 6946. *Astrophys. J.* **258**, 467–489 (1982).
9. Young, J. S. & Scoville, N. Z. Molecular gas in galaxies. *Annu. Rev. Astron. Astrophys.* **29**, 581–625 (1991).
10. Dickman, R. L., Snell, R. L. & Schloerb, F. P. Carbon monoxide as an extragalactic mass tracer. *Astrophys. J.* **309**, 326–330 (1986).
11. Downes, D. & Solomon, P. M. Rotating nuclear rings and extreme starbursts in ultraluminous galaxies. *Astrophys. J.* **507**, 615–654 (1998).
12. Smail, I., Ivison, R. J. & Blain, A. W. A deep sub-millimeter survey of lensing clusters: a new window on galaxy formation and evolution. *Astrophys. J.* **490**, L5–L8 (1997).
13. Ivison, R. J. *et al.* An excess of submillimeter sources near 4C 41.17: a candidate proto-cluster at $z = 3.8$? *Astrophys. J.* **542**, 27–34 (2000).
14. Papadopoulos, P. P. *et al.* CO(4–3) and dust emission in two powerful high- z radio galaxies, and CO lines at high redshifts. *Astrophys. J.* **528**, 626–636 (2000).
15. Plume, R. *et al.* Large-scale $^{13}\text{CO } J = 5-4$ and [CI] mapping of Orion A. *Astrophys. J.* **539**, L133–L136 (2000).
16. Aalto, S., Booth, R. S., Black, J. H. & Johansson, L. E. B. Molecular gas in starburst galaxies: line intensities and physical conditions. *Astron. Astrophys.* **300**, 369–384 (1995).
17. Jackson, J. M., Paglione, T. A. D., Carlstrom, J. E. & Rieu, N. Submillimeter HCN and HCO+ emission from galaxies. *Astrophys. J.* **438**, 695–701 (1995).
18. Paglione, T. A. D., Jackson, J. M. & Ishizuki, S. The average properties of the dense molecular gas in galaxies. *Astrophys. J.* **484**, 656–663 (1997).
19. Papadopoulos, P. P. & Allen, M. L. Gas and dust in NGC 7469: submillimeter imaging and CO $J = 3-2$. *Astrophys. J.* **537**, 631–637 (2000).
20. Hughes, D. H., Gear, W. K. & Robson, E. I. The submillimetre structure of the starburst nucleus in M82—a diffraction-limited 450-micron map. *Mon. Not. R. Astron. Soc.* **270**, 641–649 (1994).
21. Ivison, R. J., Papadopoulos, P. P., Seaquist, E. R. & Eales, S. A. A search for molecular gas in a high-redshift radio galaxy. *Mon. Not. R. Astron. Soc.* **278**, 669–672 (1996).
22. Carilli, C. L. & Holdaway, M. A. Tropospheric phase calibration in millimeter interferometry. *Radio Sci.* **34**, 817–840 (1999).
23. Carilli, C. L., Menten, K. M. & Yun, M. S. Detection of CO(2–1) and radio continuum emission from the $z = 4.4$ QSO BRI 1335–0417. *Astrophys. J.* **521**, L25–L28 (1999).

Acknowledgements

P.P. and R.I. would like to thank E. Seaquist for early discussions and encouragement. We also thank R. Barvainis. The National Radio Astronomy Observatory is a facility of the National Science Foundation, operated under cooperative agreement by Associated Universities, Inc.

Correspondence and requests for materials should be addressed to R.I. (e-mail: rji@star.ucl.ac.uk).

Substantial reservoirs of molecular hydrogen in the debris disks around young stars

W. F. Thi*, G. A. Blake†, E. F. van Dishoeck*, G. J. van Zadelhoff*, J. M. M. Horn‡, E. E. Becklin‡, V. Mannings§, A. I. Sargent||, M. E. van den Ancker¶ & A. Natta#

* Leiden Observatory, PO Box 9513, 2300 Leiden, The Netherlands

† Division of Geological and Planetary Sciences, Caltech 150-21,

§ SIRT Science Center, MS 314-6, Caltech, || Division of Physics,

Mathematics and Astronomy, Caltech 105-24, Pasadena, California 91125, USA

‡ Department of Physics and Astronomy, UCLA, California 90095-1562, USA

¶ Harvard Smithsonian Center for Astrophysics, 60 Garden Street, MS 42, Cambridge, Massachusetts 02138, USA

Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, I-50125 Firenze, Italy

Circumstellar accretion disks transfer matter from molecular clouds to young stars and to the sites of planet formation. The disks observed around pre-main-sequence stars have properties consistent with those expected for the pre-solar nebula from which our own Solar System formed 4.5 Gyr ago¹. But the ‘debris’ disks that encircle more than 15% of nearby main-sequence stars^{2–5} appear to have very small amounts of gas, based on observations of the tracer molecule carbon monoxide^{6–8}:

these observations have yielded gas/dust ratios much less than 0.1, whereas the interstellar value is about 100 (ref. 9). Here we report observations of the lowest rotational transitions of molecular hydrogen (H_2) that reveal large quantities of gas in the debris disks around the stars β Pictoris, 49 Ceti and HD135344. The gas masses calculated from the data are several hundreds to a thousand times greater than those estimated from the CO observations, and yield gas/dust ratios of the same order as the interstellar value.

We used the Short Wavelength Spectrometer¹⁰ (SWS) aboard the Infrared Space Observatory¹¹ (ISO), which provided the first opportunity to conduct measurements of the H_2 $J = 2 \rightarrow 0$ S(0) 28 μm and $J = 3 \rightarrow 1$ S(1) 17 μm pure rotational transitions above the Earth's atmosphere. These lines are the lowest energy transitions of the H_2 molecule, but are difficult (17 μm) to impossible (28 μm) to observe from the ground because of telluric absorption. A number of classical T Tauri stars, Herbig Ae stars, and young (pre-)main-sequence stars with confirmed circumstellar disks were observed. The results of the searches toward T Tauri and Herbig Ae stars may be found elsewhere^{12,13}. Here we focus on three sources previously classified as 'Vega type' objects; that is, low-mass disks with a dust population that is assumed to be dominated by second-generation grains³.

High-sensitivity mid-infrared observations of H_2 such as those made possible by ISO have a number of advantages over more commonly used indirect tracers for assessing the molecular gas content of debris disks. For a gas of solar composition, molecular hydrogen is the overwhelmingly most abundant gaseous species (by four orders of magnitude), obviating the uncertainties associated with the conversion factors needed for other molecules.

The homonuclear nature of H_2 makes the intrinsic pure rotational transition strengths sufficiently small that the optical depths of the mid-infrared lines are low, greatly simplifying the radiative transfer analysis. Because the lines appear in emission and are optically thin, any disk orientation can be studied. Substantially higher temperatures are required for excitation of the H_2 vibration-rotation manifold near 2 μm , and special geometries are needed to conduct either near-infrared or vacuum ultraviolet (electronic) absorption-line measurements. Thus, observations of the H_2 mid-infrared lines are the most direct way of probing the bulk of the warm molecular gas in the inner ~ 100 AU of disks.

The observed intensities of the H_2 lines are close to the ISO sensitivity limit, as Fig. 1 shows, and so required the development of special software to supplement the standard SWS Interactive Analysis Package. A detailed description of the software used to improve the SWS dynamic range can be found elsewhere^{12,14} (see Fig. 1 legend for details of the observing procedure). Table 1 presents the integrated line fluxes observed toward β Pictoris, 49 Ceti, and HD135344. The S(1)/S(0) ratios yield a direct measure of the excitation temperature provided the gas densities are large enough that collisional rates exceed ultraviolet pumping and spontaneous emission from the $J = 2$ and 3 levels ($n_{\text{cr}} \geq 10^4 \text{ cm}^{-3}$). For both the debris disks reported here and the disks surrounding T Tauri and Herbig Ae stars¹², temperatures near 100 K are found—values comparable to the dust temperatures derived from fits of debris disk spectral energy distributions¹⁵. The non-detection of the $J = 4 \rightarrow 2$ S(2) and $J = 5 \rightarrow 3$ S(3) lines at 12.278 and 9.662 μm limits the gas temperature to less than 250 K.

The total gas mass is very sensitive to temperature because the transition upper states lie at 510 and 1,015 K above ground, respectively, but, as outlined in Table 1, is straightforward to derive if thermal equilibrium is assumed and the distance to the source is known. Values range from a minimum of $0.17 M_J$ (Jupiter mass) for β Pictoris to just over $6.6 M_J$ for HD135344. Only an upper limit to the S(1) flux is available for 49 Ceti, which constrains the temperature at less than 100 K, corresponding to gas masses of at least $0.35 M_J$. The uncertainties in the inferred masses are estimated to be a factor of three, and stem mainly from the weakness of the lines and the unknown molecular hydrogen ortho/para ratio in such disks.

The line shapes are unresolved at the velocity resolution available to ISO, and 49 Ceti and HD135344 are too far away for their disks to be resolved with the large beams sampled by the SWS. Although β Pictoris is extended in the ISO-SWS beam, only observations centred on the star have been performed. As a result, no spatial information is available for the gas, and the fluxes obtained thus yield only lower limits for the gas mass.

Table 1 also lists the minimum dust masses derived from ISO measurements of the 60–240 μm continuum fluxes by assuming the emitting particles are small compared to the observing wavelength¹⁵ along with the gas mass estimated by CO observations. As can be seen, the gas mass exceeds that inferred for dust grains by

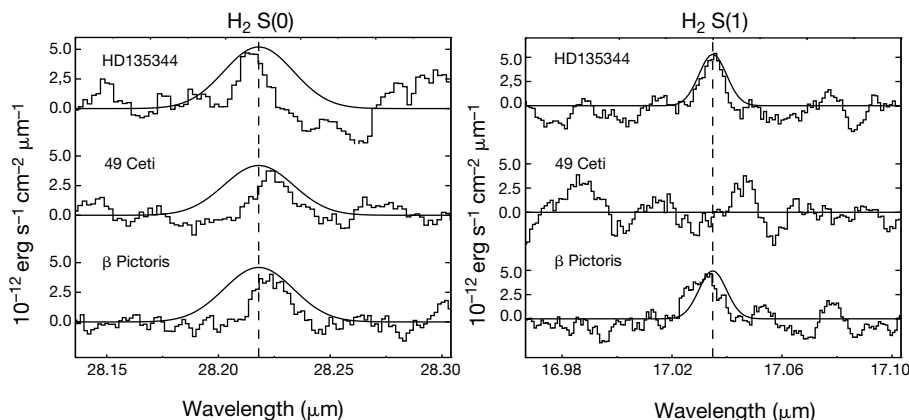


Figure 1 Spectra of the lowest pure rotational molecular hydrogen transitions toward the β Pictoris, 49 Ceti, and HD135344 debris disks taken with the Infrared Space Observatory. The SWS mode 2 was used¹⁰, with an aperture of $20'' \times 27''$ at the S(0) line and $14'' \times 27''$ at the S(1) transition, corresponding to $400 \text{ AU} \times 540 \text{ AU}$ and $280 \text{ AU} \times 540 \text{ AU}$, respectively, at the distance of β Pictoris ($1 \text{ AU} = \text{Sun-Earth distance} = 1.5 \times 10^{13} \text{ cm}$). The dashed vertical line depicts the rest wavelengths of the H_2 transitions, and the solid curve illustrates the range of potential wavelength shifts induced by grating repositioning errors or pointing offsets of the spacecraft. All line shapes

are consistent with the instrumental spectral resolution of $\lambda/\Delta\lambda \approx 2,000/2,400$ at $28.218/17.035 \mu\text{m}$. Typical integration times were 600–1,000 s per line, in which the 12 SWS detectors were scanned several times over the $28.05\text{--}28.40$ and $16.96\text{--}17.11 \mu\text{m}$ intervals around the H_2 lines. Individual scans are examined for cosmic ray glitches, dark current drifts, and readout non-linearities before being summed to produce a final spectrum^{12,14}. Off-source integrations toward 49 Ceti and HD135344 reveal no H_2 S(1) emission to a limiting flux of $8 \times 10^{-15} \text{ ergs s}^{-1} \text{ cm}^{-2}$ root mean square.

approximately the interstellar gas-to-dust ratio⁹ of 100, but the total mass is below that of the so-called minimum mass solar nebula required to generate our own planetary system, $\sim 10M_J$. The H_2 gas provides a reservoir from which gaseous giant planets may form, but, except for HD135344, the gas masses are insufficient to form any new Jupiter-mass planets in these optically thin disks. Indeed, provided the ages of these systems can be well constrained, such molecular hydrogen observations provide direct upper bounds for the formation timescales of jovian planets.

The presence of H_2 but absence of CO can be understood from models of the photodestruction of these molecules in disks¹⁶. The ultraviolet radiation from the star and from the ambient interstellar radiation field can destroy CO in tenuous disks and in the outer layers of more massive disks, whereas freeze-out of the molecule onto the surfaces of grains¹⁷ occurs in the cold inner part of massive disks. In contrast, H_2 is not significantly removed by photolysis until the disk mass falls below $10^{-3}M_J$, and it does not freeze out onto grains. Millimetre-wave studies of both 49 Ceti and HD135344 have detected weak CO rotational line emission^{7,12} consistent with gas gravitationally bound to the star, but the inferred molecular gas abundances are factors of 100–1,000 lower than found here. This has led to the perception that these and other ‘debris’ disks are largely free of gas, but our H_2 observations indicate that this is not the case. The mechanisms through which the original massive disks were dissipated remain uncertain. Presumably, this involves a combination of accretion of matter inward onto the star and removal through interaction with a stellar wind and photo-evaporation, if the ultraviolet radiation from the star is sufficiently powerful¹⁸.

It has been postulated that various transitory optical features observed toward β Pictoris result from the evaporation of infalling planetesimals^{19,20}, induced by dynamically important companions. It is difficult, however, to envision the required H_2 mass being generated by secondary processes after disk dissipation. Furthermore, the high gas/dust ratios can significantly alter the dust transport dynamics. For disks with a negative radial pressure gradient, gas surface densities some 100–1,000 times lower than those originally present enlarge the grain size for which rapid outward drift occurs to diameters of $\leq 100 \mu m$, owing to the non-Keplerian velocity field of the gas and the co-rotation of the gas and dust²¹. Larger particles drift inward. If the gas mass is sufficiently large, dust generation by collisional cascade can be shut off completely even in optically thin disks²².

For β Pictoris, the low gas mass enhances the need for the *in situ* generation of 1–100 μm dust grains²³, because such particles can be efficiently ejected from the system; the gas/dust ratio of ~ 100 found for β Pictoris could thus be coincidental in that substantial amounts of dust and gas may be hidden in large bodies. The gas masses derived for HD135344 and, perhaps, 49 Ceti may be large enough to prevent a collisional cascade of dust generation, yet increase the importance of radiative dust ejection, provided the disks are

optically thin. As such, they may be better interpreted as tenuous remnants of substantially more massive accretion disks. If so, HD135344 and 49 Ceti could be extremely interesting transition objects in which jovian planet formation may no longer be possible but for which the residual gas allows us to image gravitational perturbations in the disks owing to existing planets formed at earlier stages.

Central to this argument are the likely ages of the star/disk systems. For early-type field stars, the isochrones that are used to estimate ages based on stellar properties such as luminosity and effective temperature are consistent with two solutions. Vega, for example, can be interpreted either as a pre-main-sequence object a few Myr old or a main-sequence star whose age is ≥ 300 Myr. In most debris-disk studies, it is this older age that has been assumed^{4,5}. The older ages lead to the natural conclusion that the dust must be regenerated *in situ*, because radiation pressure and Poynting–Robertson drag forces provoke the loss of less than micrometre-sized grains within only 0.1–1 Myr in gas-free disks, and also that dust disks are the most observable features of planetary systems like our own.

The age of β Pictoris has recently been estimated to be 20 ± 10 Myr using nearby M dwarfs argued to be part of the same loose cluster on the basis of proper motion studies and *Hipparcos* distances²⁴. Ages for 49 Ceti and HD135344 derived by recent pre-main-sequence stellar evolution models²⁵ are 8 and 17 Myrs, respectively, as shown in Table 1. In addition, HD135344 possesses certain features reminiscent of young Herbig Ae stars²⁶. Thus, these objects may in fact be the descendants of circumstellar disks formed in a recent episode of local star formation.

Older stellar ages are consistent with the prevailing view that the probability of stars with ages of less than 10–30 Myr lying within a few tens of parsecs to the Sun is unlikely. The discovery of four nearby ($d < 50$ –60 pc) young associations suggests, however, that local star formation happened as recently as 10–50 Myr ago, and that the Sun now lies within this stream of young stars²⁷. For example, the G2 star/disk system HD207129 lies at a similar distance, as does β Pictoris, and has alternatively been assumed to be either 4.7 Gyr old²⁸ or only a few tens of millions of years in age²⁷. Indeed, the sensitivity limitations of the IRAS all-sky survey that is the basis for most debris-disk searches and the ISO instruments used for follow-up ensured that only nearby or fairly massive disks have been located until now^{3,29}. Searches for H_2 such as that performed here can help to establish the gas/dust mass and evolutionary state of the sources, but only a small number of objects have been surveyed so far with adequate sensitivity. Our current understanding of remnant and debris disks may therefore be biased by the local star formation history and instrumental sensitivity. The present results suggest that jovian planet formation can occur on timescales of up to 20 Myr, but we stress the limited nature of the present sample. More massive disks easily capable of jovian planet formation are seen around several T Tauri and Herbig Ae stars of

Table 1 Infrared Space Observatory H_2 observations of debris disks

Source	d (pc)	Age (Myr)	H_2 S(0)	H_2 S(1)	T_{ex} (K)	H_2 mass (M_J)	Dust mass ($10^{-2} M_J$)	M_{H_2} via CO (M_J)
β Pictoris	19.3	20 ± 10	7.0	7.7	109	0.17	0.3	$< 6 \times 10^{-5}$
49 Ceti	61	8 ± 4	6.6	< 3	< 100	> 0.35	0.63	0.006
HD135344	80	17 ± 3	9.0	5.5	97	6.6	8.5	0.005

The integrated molecular hydrogen rotational line intensities, gas masses and dust masses toward the β Pictoris, 49 Ceti, and HD135344 debris disks. The integrated intensities of the $H_2 J = 2 \rightarrow 0$ [S(0)] and $J = 3 \rightarrow 1$ [S(1)] lines at 28.218 and 17.035 μm are presented in units of 10^{-14} ergs s^{-1} cm^{-2} . Masses (in M_J) are derived using local thermodynamic equilibrium (LTE) from the expression

$$M_{H_2} = 1.76 \cdot 10^{-17} \frac{F_{ul} d^2}{(hc/4\pi\lambda) A_{ul} X_u}$$

where d is the distance in parsec, F_{ul} is the flux of either the S(0) or S(1) H_2 transition, h is Planck’s constant, c is the speed of light, λ the observing wavelength, and A_{ul} is the Einstein spontaneous emission coefficient. The upper state population x_u for the line in question is very sensitive to the excitation temperature, which is fitted from the S(1)/S(0) ratio assuming that the ortho/para ratio is in LTE at the excitation temperature T_{ex} (ortho/para = 1.6 for $T_{\text{ex}} = 100$ K). Numerically, $T_{\text{ex}} = 505.24 \ln[1.125 S(0)/S(1)]$, in K. As upper limits only are available for the S(1) flux from 49 Ceti, the H_2 mass is calculated for the upper temperature limit of 100 K, resulting in a lower bound for the mass. Dust masses are from the ISO measurements of ref. 15. The H_2 mass for β Pictoris refers to the inner 400 A_U , which contains at least 80% of the disk mass¹⁶. The age estimates for 49 Ceti and HD135344 are derived using the pre-main-sequence stellar evolution tracks of ref. 25, and that for β Pictoris is taken from ref. 24. The main-sequence upper limit to the age of β Pictoris is approximately 100 Myr; the limits for 49 Ceti and HD135344 are even younger.

only slightly younger ages, for example, and illustrate the pressing need for better statistics¹².

The Space Infrared Telescope Facility (SIRTF) to be launched by NASA in 2002 will go much deeper than IRAS and ISO, and will provide hundreds of new targets for further study. Improved molecular-line observations will also become possible with the high-resolution mid-infrared spectrometers nearing completion for use on large ground-based telescopes (for the 17- μm S(1) line) and the Stratospheric Observatory for Infrared Astronomy (SOFIA) (for all H₂ lines including the 28- μm S(0) transition). These diffraction-limited spectrometers will possess both the spatial and spectral resolution required to examine whether the gas and dust are co-spatial in the disks, and can establish whether the gas persists to within a few AU of the central star without the use of coronagraphic techniques. In the longer term, a mid-infrared spectrometer on the Next Generation Space Telescope (NGST) would make possible the detection of Earth-masses of H₂ gas at temperatures down to 50 K. By combining continuum and line studies of disks around stars up to several tens of Myr in age, it will be possible to determine both gas/dust dissipation and jovian planet formation timescales; and to examine the role of collisional, radiative and gaseous processes in shaping the evolution and survival of dust clouds in planetary systems. □

Received 25 September 2000; accepted 16 November 2000.

- Beckwith, S. W. V. & Sargent, A. I. Circumstellar disks and the search for neighbouring planetary systems. *Nature* **383**, 139–144 (1996).
- Aumann, H. *et al.* Discovery of a shell around Alpha Lyrae. *Astrophys. J.* **278**, L23–L27 (1984).
- Backman, D. E. & Paresce, F. in *Protostars & Planets III* (eds Levy, E. H. & Lunine, J. I.) 1253–1304 (Univ. Arizona Press, Tucson, 1993).
- Habing, H. J. *et al.* Disappearance of stellar debris disks around main-sequence stars after 400 million years. *Nature* **401**, 456–458 (1999).
- Lagrange, A. M., Backman, D. E. & Artymowicz, P. in *Protostars & Planets IV* (eds Mannings, V., Boss, A. & Russell, S.) 639–672 (Univ. Arizona Press, Tucson, 2000).
- Greaves, J. S., Coulson, I. M. & Holland, W. S. No molecular gas around nearby solar-type stars. *Mon. Not. R. Astron. Soc.* **312**, L1–L3 (2000).
- Zuckerman, B., Forveille, T. & Kastner, J. H. Inhibition of giant planet formation by rapid gas depletion around young stars. *Nature* **373**, 494–496 (1995).
- Dent, W. R. F., Greaves, J. S., Mannings, V., Coulson, I. M. & Walther, D. M. A search for molecular gas components in prototypical Vega-excess systems. *Mon. Not. R. Astron. Soc.* **277**, L25–L29 (1995).
- Hildebrand, R. H. The determination of cloud masses and dust characteristics from submillimeter thermal emission. *Q. J. R. Astron. Soc.* **24**, 267–282 (1983).
- de Graauw, T. *et al.* Observing with the ISO Short Wavelength Spectrometer. *Astron. Astrophys.* **315**, L49–L54 (1996).
- Kessler, M. *et al.* The Infrared Space Observatory (ISO) mission. *Astron. Astrophys.* **315**, L64–L70 (1996).
- Thi, W. F. *et al.* H₂ and CO emission from disks around T Tauri, Herbig Ae and Vega-type stars: cold and warm circumstellar gas. *Astrophys. J.* (submitted).
- Thi, W. F., van Dishoeck, E. F., Blake, G. A., van Zadelhoff, G. J. & Hogerheijde, M. R. Detection of H₂ pure rotational line emission from the GG Tau circumbinary disk. *Astrophys. J.* **521**, L63–L66 (1999).
- Valentijn, E. A. & Thi, W. F. ISO's Short Wavelength Spectrometer—Ultimate sensitivity. Reducing the effects of cosmic weather. *Exp. Astron.* **10**, 215–225 (2000).
- Walker, H. J. & Heinrichsen, I. ISOPHOT observations of dust disks around main-sequence (Vega-like) stars. *Icarus* **143**, 147–154 (2000).
- Kamp, I. & Bertoldi, F. CO in the circumstellar disks of Vega and β Pictoris. *Astron. Astrophys.* **353**, 276–286 (2000).
- Sandford, S. A. & Allamandola, L. J. H₂ in interstellar and extragalactic ices—Infrared characteristics, ultraviolet production, and implications. *Astrophys. J.* **409**, L65–L68 (1993).
- Hollenbach, D. J., Yorke, H. W. & Johnstone, D. in *Protostars & Planets IV* (eds Mannings, V., Boss, A. & Russell, S.) 401–428 (Univ. Arizona Press, Tucson, 2000).
- Lagrange, A. M. *et al.* The β Pic circumstellar disk. XXIV. Clues to the origin of the stable gas. *Astron. Astrophys.* **310**, 1091–1108 (1998).
- Heap, S. R. *et al.* STIS coronagraphic observations of β Pictoris. *Astrophys. J.* **539**, 435–444 (2000).
- Klahr, H. H. & Lin, D. N. C. Dust distribution in gas disks. A model for the ring around HR 4796A. *Astron. J.* (in the press).
- Weidenschilling, S. J. Aerodynamics of solid bodies in the solar nebula. *Mon. Not. R. Astron. Soc.* **180**, 57–70 (1977).
- Artymowicz, P. & Clampin, M. Dust around main sequence stars: Nature or nurture by the interstellar medium? *Astrophys. J.* **490**, 863–878 (1997).
- Barrado y Navascués, D., Stauffer, J. R., Song, I. & Caillaud, J. P. The age of β Pictoris. *Astrophys. J.* **520**, L123–L126 (1999).
- Siess, L., Forestini, M. & Bertout, C. Physics of accretion onto young stars. III. Comparisons with observations. *Astron. Astrophys.* **342**, 480–491 (1999).
- Dunkin, S. K., Barlow, M. J. & Ryan, S. G. High-resolution spectroscopy of Vega-like stars—II. Age indicators, activity and circumstellar gas. *Mon. Not. R. Astron. Soc.* **290**, 165–185 (1997).
- Zuckerman, B. & Webb, R. A. Identification of a nearby stellar association with the *Hipparcos* catalog: Implications for recent, local star formation. *Astrophys. J.* **535**, 959–964 (2000).
- Jourdain de Muizon, M., Laureijs, R. J., Dominik, C. & Habing, H. J. A very cold disk of dust around the G2V star HD 207129. *Astron. Astrophys.* **350**, 875–882 (1999).

29. Fajardo-Acosta, S. B., Stencil, R. E., Backman, D. E. & Thakur, N. Infrared Space Observatory photometric search of main-sequence stars for Vega-type Systems. *Astrophys. J.* **520**, 215–222 (1999).

Acknowledgements

This work was supported by The Netherlands Organization for Scientific Research (NWO). Additional support to G.A.B. from the NASA Infrared Space Observatory (ISO), Exobiology, and Origins of Solar Systems programmes is gratefully acknowledged. A.N. is supported in part by an Agenzia Spaziale Italiana (ASI) grant. This work is based on observations with ISO, a European Space Agency (ESA) project with instruments funded by ESA Member States (especially the Principal Investigator countries: France, Germany, the Netherlands, and the United Kingdom) and with the participation of the Institute of Space and Astronautical Sciences (ISAS) and the National Aeronautics and Space Administration (NASA).

Correspondence should be addressed to G.A.B. (e-mail: gab@gps.caltech.edu).

Many-particle entanglement with Bose–Einstein condensates

A. Sørensen*, L.-M. Duan†, J. I. Cirac† & P. Zoller†

* Institute of Physics and Astronomy, University of Aarhus, DK-8000 Århus C, Denmark

† Institute for Theoretical Physics, University of Innsbruck, A-6020 Innsbruck, Austria

The possibility of creating and manipulating entangled states of systems of many particles is of significant interest for quantum information processing; such a capability could lead to new applications that rely on the basic principles of quantum mechanics¹. So far, up to four atoms have been entangled in a controlled way^{2,3}. A crucial requirement for the production of entangled states is that they can be considered pure at the single-particle level. Bose–Einstein condensates^{4–6} fulfil this requirement; hence it is natural to investigate whether they can also be used in some applications of quantum information. Here we propose a method to achieve substantial entanglement of a large number of atoms in a Bose–Einstein condensate. A single resonant laser pulse is applied to all the atoms in the condensate, which is then allowed to evolve freely; in this latter stage, collisional interactions produce entanglement between the atoms. The technique should be realizable with present technology.

Consider a set of N two-level atoms confined by some external trap. In order to describe the internal properties of these atoms, it is convenient to let the internal states $|a\rangle_n$ and $|b\rangle_n$ of the n th atom represent the two states of a fictitious spin-1/2 particle with angular momentum operators $j_z^{(n)} = 1/2(|a\rangle\langle a|_n - |b\rangle\langle b|_n)$, $j_x^{(n)} = 1/2(|b\rangle\langle a|_n + |a\rangle\langle b|_n)$, and $j_y^{(n)} = i/2(|b\rangle\langle a|_n - |a\rangle\langle b|_n)$. We consider collective effects of the atoms that are described by total angular momentum operators, $\mathbf{J} = \sum_{n=1}^N \mathbf{j}^{(n)}$. The entanglement properties of the atoms can be expressed in terms of the variances and expectation values of these operators. In the Methods section we show that if

$$\xi^2 \equiv \frac{N(\Delta J_{n_1})^2}{\langle J_{n_2} \rangle^2 + \langle J_{n_3} \rangle^2} < 1 \quad (1)$$

where $\mathbf{J}_n \equiv \mathbf{n} \cdot \mathbf{J}$ and the $\mathbf{n}$ s are mutually orthogonal unit vectors, then the state of the atoms is non-separable (that is, entangled). The parameter ξ^2 thus characterizes the atomic entanglement, and states with $\xi^2 < 1$ are often referred to as “spin squeezed states”⁷. Here we show how to reduce ξ^2 by several orders of magnitude using the collisional interactions between atoms in a Bose–Einstein condensate.

We consider a two-component weakly interacting Bose–Einstein condensate, which has been produced in several laboratories^{8,9}, and we assume that the interactions do not change the internal state of the atoms. This situation is described by the second quantized hamiltonian

$$H = \sum_{j=a,b} \int d^3r \hat{\Psi}_j^\dagger(\mathbf{r}) H_{0,j} \hat{\Psi}_j(\mathbf{r}) + \frac{1}{2} \sum_{j=a,b} U_{jj} \int d^3r \hat{\Psi}_j^\dagger(\mathbf{r}) \hat{\Psi}_j^\dagger(\mathbf{r}) \hat{\Psi}_j(\mathbf{r}) \hat{\Psi}_j(\mathbf{r}) \\ + U_{ab} \int d^3r \hat{\Psi}_a^\dagger(\mathbf{r}) \hat{\Psi}_b^\dagger(\mathbf{r}) \hat{\Psi}_a(\mathbf{r}) \hat{\Psi}_b(\mathbf{r}) \quad (2)$$

where $H_{0,j}$ is the one-particle hamiltonian for atoms in state j including the kinetic energy and the external trapping potential $V_j(\mathbf{r})$, $\hat{\Psi}_j(\mathbf{r})$ is the field operator for atoms in the state j , $U_{jk} = 4\pi\hbar^2 a_{jk}/m$ is the strength of the interaction between particles of type j and k , parametrized by the scattering length a_{jk} , and m is the atomic mass.

Assume that we start with a Bose–Einstein condensate in state $|a\rangle$ at very low temperature ($T \approx 0$), so that all the atoms are in a single-particle (motional) state $|\phi_0\rangle$. A fast $\pi/2$ pulse between the states $|a\rangle$ and $|b\rangle$ prepares the atoms in the state $|\phi_0\rangle^{\otimes N} \otimes (|a\rangle + |b\rangle)^{\otimes N} 2^{N/2}$ which is an eigenstate of the J_x operator with eigenvalue $N/2$. If we choose in equation (1) $\mathbf{n}_1 = [0, \cos(\theta), \sin(\theta)]$ and $\mathbf{n}_2$ along the x axis of the fictitious angular momentum, we have $\xi_\theta^2 = 1$ at $t = 0$. By using the equations of motion of the angular momentum operators in the Heisenberg picture, we find the time derivative of ξ_θ^2 at $t = 0$:

$$\frac{d}{dt} \xi_\theta^2 = \sin(2\theta) \frac{(N-1)(U_{aa} + U_{bb} - 2U_{ab})}{2\hbar} \int d^3r |\phi_0|^4 \quad (3)$$

This equation shows that spin squeezing will be produced for certain angles θ if $U_{aa} + U_{bb} \neq 2U_{ab}$, as it is in the sodium experiments⁹.

To quantify the amount of squeezing which may be obtained, we assume identical trapping potentials $V_a(\mathbf{r}) = V_b(\mathbf{r})$ and identical coupling constants for interaction between atoms in the same internal state $a_{aa} = a_{bb} > a_{ab}$. Physically, this could correspond to the $|F = 1, M_F = \pm 1\rangle$ hyperfine states of Na trapped in an optical dipole trap (F and M_F are the quantum numbers of the atomic angular momentum). Owing to the symmetry of these states, their

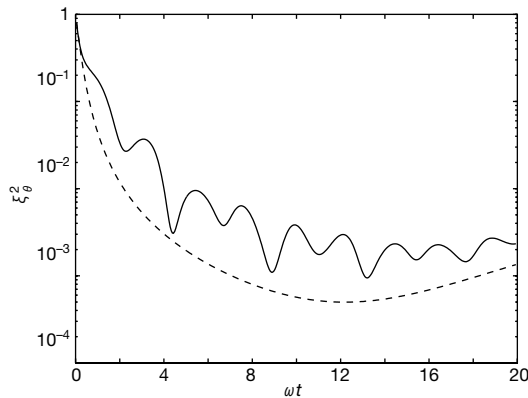


Figure 1 Reduction in the squeezing parameter ξ_θ^2 . A fast $\pi/2$ pulse between two internal states is applied to all atoms in the condensate. The subsequent free evolution with time results in a strong squeezing of the total spin. The angle θ is chosen such that ξ_θ^2 is minimal. The solid line shows the results of a numerical integration (see text). For numerical convenience we have assumed a spherically symmetric potential $V(r) = m\omega^2 r^2/2$. The parameters are $a_{aa}/d_0 = 6 \times 10^{-3}$, $a_{bb} = 2a_{ab} = a_{aa}$, and $N = 10^5$. The dashed curve shows the squeezing obtained from the hamiltonian $H_{\text{spin}} = \hbar\chi J_z^2$. The parameter χ is chosen such that the reduction of $\langle J_x \rangle$ obtained from the solution in ref. 7 is consistent with the results of ref. 11.

scattering lengths and trapping potentials will be identical; moreover, angular-momentum conservation ensures there are no spin-exchanging collisions between these states as required by our model. This is exactly the experimental set-up used in ref. 9. In this experiment it is shown that these states have an antiferromagnetic interaction $a_{ab} < a_{aa}$ which, according to equation (3), enables the production of squeezed states. To avoid spin-changing collisions that populate the state $|M_F = 0\rangle$ (ref. 10), it is necessary to modify slightly such an experimental set-up. If the $F = 1$ manifold is coupled to the $F = 2$ with a far off-resonant blue-detuned π -polarized microwave field, the $|F = 1, M_F = 0\rangle$ state is raised in energy with respect to the $|F = 1, M_F = \pm 1\rangle$ states, because the coupling coefficient for the $|F = 1, M_F = 0\rangle \rightarrow |F = 2, M_F = 0\rangle$ transition is larger than the coefficients for the $|F = 1, M_F = \pm 1\rangle \rightarrow |F = 2, M_F = \pm 1\rangle$ transitions, and spin-exchanging collisions become energetically forbidden. If, for instance, we choose a detuning $\delta = (2\pi)25$ MHz and a resonant Rabi frequency for the $1 \rightarrow 1$ transition of $\Omega = (2\pi)2$ MHz, the energy separation is $\Delta E = 640$ nK. With a typical chemical potential $\mu \approx 220$ nK (ref. 10), this energy separation is much higher than the available energy in the collisions and the $|F = 1, M = 0\rangle$ state is completely decoupled. In ref. 10 a much smaller energy difference is shown to exclude spin-exchanging collision, and we therefore expect that much weaker fields will suffice.

The assumption $a_{aa} = a_{bb}$ has several advantages. First, it reduces the effects of fluctuations in the total particle number. If $a_{aa} \neq a_{bb}$, the mean spin performs an N -dependent rotation around the z -axis, and fluctuations in the number of particles introduces an uncertainty in the direction of the spin that effectively reduces the average value and introduces noise into the system. With $a_{aa} = a_{bb}$, the mean spin remains in the x -direction, independent of the number of atoms in the trap. Second, this condition ensures a large spatial overlap of different components of the wavefunction. After the $\pi/2$ pulse, the spatial wavefunction is no longer in the equilibrium state. That is, owing to the atomic repulsions (which are now different from before because the atoms are in different internal states), the spatial distribution of the atomic cloud will start oscillating. Furthermore, as the state of the system is now distributed over a range of number of particles in the $|a\rangle$ state (N_a), and because this number enters into the time evolution, the N_a -dependent wavefunctions ϕ_a and ϕ_b are different for particles in the states $|a\rangle$ and $|b\rangle$. With $a_{aa} = a_{bb}$, ϕ_a and ϕ_b are identical if N_a equals the average number $N/2$. In the limit of large N , the width of the distribution on different N_a s is much smaller than N_a and all the spatial wavefunctions are approximately identical: $\phi_a(N_a, t) \approx \phi_b(N_a, t) \approx \phi_0(t)$. This relation is only true if $a_{aa} > a_{ab}$ where small deviations from

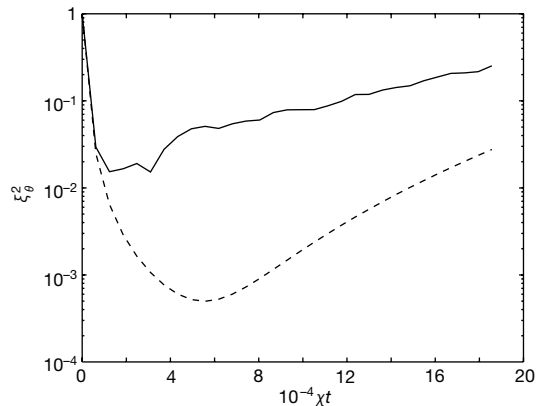


Figure 2 Quantum Monte Carlo simulation of squeezing in the presence of loss. The solid line shows squeezing obtained from the hamiltonian H_{spin} with particle loss described by a constant loss rate $\Gamma = 200\chi$. The dashed curve shows squeezing without particle loss.

the average wavefunction perform small oscillations. In the opposite case the deviations grow exponentially¹¹, resulting in a reduction of the overlap of the a and the b wavefunctions and a reduced squeezing. The advantages mentioned above could also be achieved with $a_{aa} \neq a_{bb}$ by using the breath-together solutions proposed in ref. 11.

Before analysing the complete system quantitatively, we estimate the amount of spin squeezing we can reach with our proposal by using a simple model. Assuming that the same wavefunction ϕ_0 for both $|a\rangle$ and $|b\rangle$ atoms is constant and independent of the number N_a , the spin-dependent part of the hamiltonian (equation (2)) may be written as $H_{\text{spin}} = \hbar\chi J_z^2$, where χ depends on the scattering lengths and the wavefunction ϕ_0 . The spin squeezing produced by this hamiltonian can be calculated exactly⁷. In the limit of large N , the minimum obtainable squeezing parameter is $\xi_\theta^2 = \frac{1}{2}(3/N)^{2/3}$, which indicates that our proposal might produce a reduction of ξ^2 by a factor of $\sim N^{2/3}$; this would be more than three orders of magnitude with 10^5 atoms in the condensate.

In contrast to the simplified hamiltonian H_{spin} , the real hamiltonian (equation (2)) will also entangle the internal and motional states of the atoms; this is a source of decoherence for the spin squeezing. To quantify this effect, we have performed a direct numerical integration following the procedure developed in ref. 11. We split the whole Hilbert space into orthogonal subspaces containing a fixed number of particles N_a and $N_b = N - N_a$ in each of the internal states, respectively. In each subspace we make a Hartree–Fock variational ansatz in terms of three-dimensional spatial wavefunctions $\phi_a(N_a, t)$ and $\phi_b(N_b, t)$, which are evolved according to the time-dependent coupled Gross–Pitaevskii equations⁶. This is an approximation to the full problem which is valid in the limit of low temperatures and short times, where the population of the Bogoliubov modes⁶ is small. In particular, it is a good approximation in our case with $a_{ab} < a_{aa}, a_{bb}$ where there are no demixing instabilities. With this procedure, the decoherence induced by the entanglement with the motional state is effectively taken into account. Together with the prediction from the simple hamiltonian H_{spin} , the result of the simulation is shown in Fig. 1. The two curves are roughly in agreement, confirming that the system is able to approximate the results of the hamiltonian H_{spin} . The numerical solution shows fluctuations due to the oscillations of the spatial wavefunction. The large dips at $\omega t \approx 4, 9, 13$ and 18 are the points where the atomic cloud reaches the initial width. At these instants the overlap of the wavefunctions is maximal and the two curves are very close (up to a factor of two). The small dips at $\omega t = 2, 7, 11$ and 16 correspond to the points of maximum compression. With the realistic parameters used in the figure, our simulation suggests that squeezing of three orders of magnitude is possible. Also, we note the timescale in the figure. The maximally squeezed state is reached after approximately two oscillation periods in the trap. For a fixed ratio of the scattering lengths a_{ab}/a_{aa} , the optimal time scales as $(a_{aa}/d_0)^{-2/5} N^{-1/15}$, where $d_0 = [\hbar/(m\omega)]^{1/2}$ is the width of the ground state of the harmonic potential.

The analysis so far has left out a number of possible imperfections. Specifically, we have assumed that all scattering lengths are real so that no atoms are lost from the trap, and we have not considered the role of thermal particles. To estimate the effect of particle losses, we have performed a Monte Carlo simulation¹² of the evolution of squeezing from the hamiltonian H_{spin} . The particle loss is phenomenologically taken into account by introducing a loss rate Γ which is identical for atoms in the $|a\rangle$ and the $|b\rangle$ state. In Fig. 2 we show the obtainable squeezing in the presence of loss. Approximately 10% of the atoms are lost at the time $\chi t \approx 6 \times 10^{-4}$ where the squeezing is optimal for the curve without loss. With the parameters of Fig. 1 this time corresponds to roughly two trapping periods. Such a large loss is an exaggeration of the loss compared to current experiments and the simulations indicate that even under these conditions, squeezing of nearly two orders of magnitude may

be obtained. On the other hand, the effects of thermal particles can be suppressed at sufficiently low temperatures but, owing to the robustness with respect to particle losses shown in Fig. 2, we expect to obtain high squeezing even at some finite temperatures.

We believe that we have presented a simple and robust method to produce entangled states of a large number of atoms with present-day technology. These entangled states are of fundamental physical interest, and they also have possible technological applications in atomic clocks^{13,14}, where the projection noise $(\Delta J_\theta)^2$ is currently the main source of noise¹⁵. Boyer and Kasevich¹⁴ have shown how to use an entangled state with half of the particles $N/2$ in one internal state a , and the other half in the other state b to obtain the Heisenberg limit in atomic interferometric measurements. On the other hand, it has been shown that if the atoms are prepared in a state with $\xi^2 < 1$ before they are injected into an atomic clock, it is possible to reduce the frequency noise (variance in the frequency measurements) or the measuring time to obtain a desired precision by a factor of ξ^2 compared to when the atoms are used in an uncorrelated state¹³. If the squeezing produced by our proposal was transferred into a suitable clock transition, and the atomic cloud was allowed to expand so that the role of collisions was reduced during the frequency measurement, the states could be directly applied in the current set-up of atomic clocks. Other theoretical proposals for noise reduction in systems of neutral atoms have been made¹⁶, and a weak squeezing of the spin has recently been produced experimentally^{17,18}. However, our proposal has the advantage that it offers a very strong noise reduction, and it is directly applicable to existing experimental set-ups. In future experiments that have negligible particle loss even for very long interaction times, the hamiltonian H_{spin} could also be used to produce a maximally entangled state of any number of atoms^{2,19}, which could reduce the frequency noise to the fundamental limit of quantum mechanics²⁰. □

Methods

Here we present the derivation of equation (1) as a criterion for entanglement. An N -particle density matrix ρ is defined to be separable (non-entangled) if it can be decomposed into

$$\rho = \sum_k P_k \rho_k^{(1)} \otimes \rho_k^{(2)} \otimes \dots \otimes \rho_k^{(N)} \quad (4)$$

where the coefficients P_k are positive real numbers fulfilling $\sum_k P_k = 1$, and $\rho_k^{(i)}$ is a density matrix for the i th particle. The variance of J_z may be described as $(\Delta J_z)^2 = (N/4) - \sum_i P_i \Sigma_i \langle j_z^{(i)} \rangle^2 + \sum_i P_i \langle j_z^{(i)} \rangle^2 - \langle J_z \rangle^2$, and using Cauchy–Schwarz’s inequality and $\langle j_z^{(i)} \rangle^2 + \langle j_y^{(i)} \rangle^2 + \langle j_x^{(i)} \rangle^2 \leq 1/4$ we find three inequalities for separable states $\Sigma_i P_i \langle j_z^{(i)} \rangle^2 \geq \langle J_z \rangle^2$, $-\sum_i P_i \Sigma_i \langle j_z^{(i)} \rangle^2 \geq -(N/4) + \sum_i P_i \Sigma_i \langle j_z^{(i)} \rangle^2 + \langle j_z^{(i)} \rangle^2$, and $\langle J_z \rangle^2 \leq N \sum_i P_i \Sigma_i \langle j_z^{(i)} \rangle^2$. From these inequalities we find that any separable state obeys $\xi^2 \geq 1$ and hence any state with $\xi^2 < 1$ is non-separable.

Received 26 June; accepted 10 November 2000.

- Special issue on quantum information. *Phys. World* **11** (2) 33–57 (1998).
- Sackett, C. A. *et al.* Experimental entanglement of four particles. *Nature* **404**, 256–259 (2000).
- Rauschenbeutel, A. *et al.* Step-by-step engineered multiparticle entanglement. *Science* **288**, 2024–2028 (2000).
- Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E. & Cornell, E. A. Observation of Bose–Einstein condensation in a dilute atomic vapor. *Science* **269**, 198–201 (1995).
- Davis, K. B. *et al.* Bose–Einstein condensation in a gas of sodium atoms. *Phys. Rev. Lett.* **75**, 3969–3973 (1995).
- Parkins, A. S. & Walls, D. F. The physics of trapped dilute-gas Bose–Einstein condensates. *Phys. Rep.* **303**, 1–80 (1998).
- Kitagawa, M. & Ueda, M. Squeezed spin states. *Phys. Rev. A* **47**, 5138–5143 (1993).
- Hall, D. S., Matthews, M. R., Ensher, J. R., Wieman, C. E. & Cornell, E. A. Dynamics of component separation in a binary mixture of Bose–Einstein condensates. *Phys. Rev. Lett.* **81**, 1539–1542 (1998).
- Stenger, J. *et al.* Spin domains in ground-state Bose–Einstein condensates. *Nature* **396**, 345–348 (1998).
- Miesner, H.-J. *et al.* Observation of metastable states in spinor Bose–Einstein condensates. *Phys. Rev. Lett.* **82**, 2228–2231 (1999).
- Sinatra, A. & Castin, Y. Binary mixtures of Bose–Einstein condensates: Phase dynamics and spatial dynamics. *Eur. Phys. J. D* **8**, 319–332 (2000).
- Molmer, K., Castin, Y. & Dalibard, J. Monte Carlo wave-function method in quantum optics. *J. Opt. Soc. Am. B* **10**, 524–538 (1993).
- Wineland, D. J., Bollinger, J. J., Itano, W. M. & Heinzen, D. J. Squeezed atomic states and projection noise in spectroscopy. *Phys. Rev. A* **50**, 67–88 (1994).

14. Boyer, P. & Kasevich, M. A. Heisenberg-limited spectroscopy with degenerate Bose-Einstein gases. *Phys. Rev. A* **56**, R1083–R1086 (1997).
15. Santarelli, G. *et al.* Quantum projection noise in an atomic fountain: A high stability cesium frequency standard. *Phys. Rev. Lett.* **82**, 4619–4622 (1999).
16. Sørensen, A. & Mølmer, K. Spin-spin interaction and spin squeezing in an optical lattice. *Phys. Rev. Lett.* **83**, 2274–2277 (1999).
17. Hald, J., Sørensen, J. L., Schori, C. & Polzik, E. S. Spin squeezed atoms: A macroscopic entangled ensemble created by light. *Phys. Rev. Lett.* **83**, 1319–1322 (1999).
18. Kuzmich, A., Mandel, L. & Bigelow, N. P. Generation of spin squeezing via continuous quantum nondemolition measurement. *Phys. Rev. Lett.* **85**, 1594–1597 (2000).
19. Mølmer, K. & Sørensen, A. Multiparticle entanglement of hot trapped ions. *Phys. Rev. Lett.* **82**, 1835–1838 (1999).
20. Bollinger, J. J., Itano, W. M., Wineland, D. J. & Heinzen, D. J. Optimal frequency measurements with maximally correlated states. *Phys. Rev. A* **54**, 4649–4652 (1996).

Acknowledgements

This work was supported by the Austrian Science Foundation, the European Union project EQUIP, the TMR European network, the ESF under the PESC program “Quantum Information”, the Institute for Quantum Information GmbH, and the Thomas B. Thriges Center for Kvantoinformatik. A.S. acknowledges the hospitality of the University of Innsbruck.

Correspondence and requests for materials should be addressed to A.S. (e-mail: anderss@ifa.au.dk).

Indium phosphide nanowires as building blocks for nanoscale electronic and optoelectronic devices

Xiangfeng Duan^{*†}, Yu Huang^{*†}, Yi Cui^{*}, Jianfang Wang^{*} & Charles M. Lieber^{*‡}

^{*} Department of Chemistry and Chemical Biology, [‡] Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

[†] These authors contributed equally to this work

Nanowires and nanotubes carry charge and excitons efficiently, and are therefore potentially ideal building blocks for nanoscale electronics and optoelectronics^{1,2}. Carbon nanotubes have already been exploited in devices such as field-effect^{3,4} and single-electron^{5,6} transistors, but the practical utility of nanotube components for building electronic circuits is limited, as it is not yet possible to selectively grow semiconducting or metallic nanotubes^{7,8}. Here we report the assembly of functional nanoscale devices from indium phosphide nanowires, the electrical properties of which are controlled by selective doping. Gate-voltage-dependent transport measurements demonstrate that the nanowires can be predictably synthesized as either n- or p-type. These doped nanowires function as nanoscale field-effect transistors, and can be assembled into crossed-wire p–n junctions that exhibit rectifying behaviour. Significantly, the p–n junctions emit light strongly and are perhaps the smallest light-emitting diodes that have yet been made. Finally, we show that electric-field-directed assembly can be used to create highly integrated device arrays from nanowire building blocks.

We prepared single-crystal InP nanowires with n- and p-type doping by laser-assisted catalytic growth (LCG)^{9,10}. Field-emission scanning electron microscopy (FE-SEM) images of the doped nanowires (Fig. 1a) show that the wires are up to tens of micrometres long, with diameters of the order of 10 nm. High-resolution transmission electron microscopy (TEM) images (Fig. 1a inset) show that the doped nanowires are single crystals with $\langle 111 \rangle$ growth directions. To confirm the presence and type of dopants in the nanowires, we have performed gate-dependent, two-terminal transport measurements on individual wires, as the conductance will

respond in an opposite way to changes in gate voltage (V_g) for n- and p-type nanowires: $V_g > 0$ will lead to an accumulation of electrons and increase in conductance for the former, but will deplete holes and reduce conductance for the latter¹¹. Figures 1b and c show the typical gate-dependent current–voltage (I – V) curves obtained from individual Te- and Zn-doped nanowires, respectively. These curves are linear or nearly linear at $V_g = 0$, indicating that the metal electrodes make ohmic contact to the nanowires. This ohmic nature was substantiated by conductance and four-terminal measurements (see Supplementary Information). The transport data recorded on Te-doped nanowires (Fig. 1b) show an increase in conductance for $V_g > 0$, while the conductance decreases for $V_g < 0$. These data clearly show that Te-doped InP nanowires are n-type. Gate-dependent transport data recorded on Zn-doped nanowires show

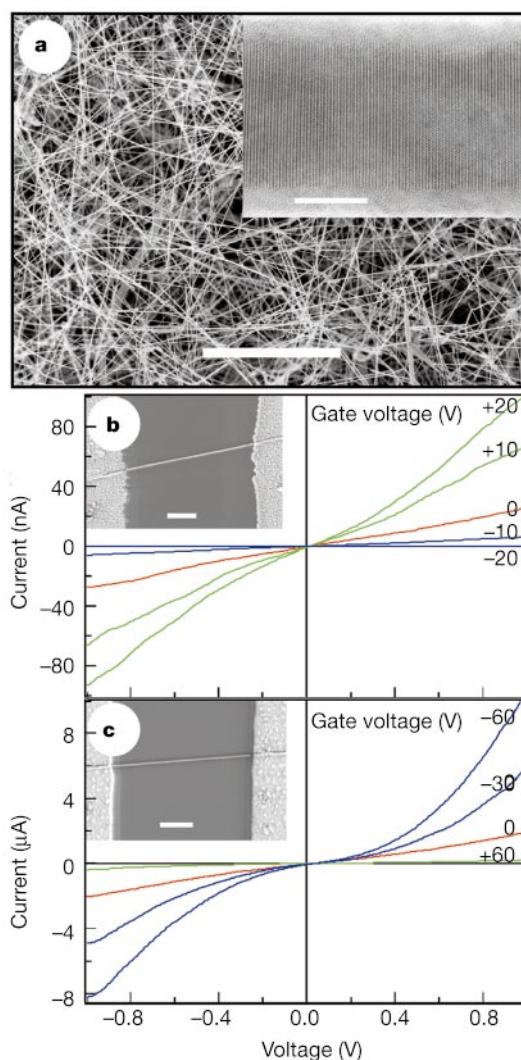


Figure 1 Doping and electrical transport of InP nanowires. **a**, A typical FE-SEM image of Zn-doped InP nanowires. Scale bar, 10 μm. Inset, lattice resolved TEM image of one 26-nm-diameter NW. The $\langle 111 \rangle$ lattice planes are visible perpendicular to the wire axis. Scale bar, 10 nm. The TEM studies show that nanowires are typically coated with a 1–2-nm amorphous overlayer. This thin layer is attributed to oxides formed when the nanowires are exposed to air after synthesis. The overall composition of individual nanowires determined by energy dispersive X-ray analysis was found to be 1:1 In:P, thus confirming their stoichiometric composition. **b** and **c**, Gate-dependent I – V behaviour for Te- and Zn-doped NWs, respectively. Insets show the nanowire measured with two-terminal Ni/In/Au contact electrodes. Scale bars, 1 μm. The diameter of the nanowire in **b** is 47 nm, while that in **c** is 45 nm. Gate voltages used in the measurements are indicated on the corresponding I – V curves (right side). Data were recorded at room temperature.

opposite changes in conductance with variation in V_g : for $V_g > 0$, conductance decreases and for $V_g < 0$ conductance increases (Fig. 1c). Zn-doped InP nanowires are therefore p-type.

Our results are reproducible. Measurements made on more than 20 individual nanowires of each dopant type show gate effects in each case that are consistent with the dopant used during synthesis. In addition, V_g can be used to completely deplete electrons and holes in n- and p-type nanowires such that the conductance becomes immeasurably small. For example, the conductance of the nanowire in Fig. 1b can be switched from a conducting (on) to an insulating (off) state when $V_g \leq -20$ V; thus it functions as a field-effect transistor (FET). Conductance modulations as large as 4–5 orders of magnitude have been observed. The relatively large switching voltage is related to the thick (600 nm) oxide dielectric used in our measurements. This gate-dependent behaviour is similar to that of metal-oxide-semiconductor (MOS) FETs¹² and recent semiconducting nanotube FETs^{3,4}. An important distinction of our work with respect to nanotubes is that predictable semi-conducting behaviour can be achieved in every nanowire. Because these InP nanowires are produced in bulk quantities, they represent a readily available material for assembling devices and device arrays.

The availability of well-defined n- and p-type nanowire building blocks opens up the possibility of creating complex functional devices by forming junctions between two or more such wires. To explore this opportunity, we have investigated n–n, p–p and p–n junctions formed by crossing two n-type, two p-type, and one n-type and one p-type nanowire, respectively. Significantly, the types of junctions studied are controllable for every experiment as we can select the types of nanowires used to produce the crossed junction before assembly. Figure 2a shows a representative crossed nanowire device formed from one of 29 nm diameter and one of 40 nm diameter.

Figures 2b and c show I – V data recorded on n–n and p–p junctions, respectively. The transport data recorded on the individual nanowires (AC, BD) show linear or nearly linear I – V behaviour (green and blue curves in Fig. 2b and c respectively). As discussed above, these results show that the metal electrodes make ohmic or

nearly ohmic contact to the nanowires, and hence will not make significant contributions to the I – V measurements across junctions. In general, transport measurements made across the n–n and p–p junctions also show linear or nearly linear behaviour, and allow us to infer two important points about crossed nanowire junctions. First, interface oxide between individual nanowires does not produce a significant tunnelling barrier, as such a barrier would lead to highly nonlinear I – V curves. Second, the I – V curves recorded through each pair (AB, AD, CB, CD in Fig. 2) of adjacent arms are similar and $I(V)$ is smaller than through individual nanowires. These results demonstrate that the junction dominates the transport behaviour. Overall, our data indicate that crossed nanowires make reasonably good electrical contact with each other, despite the small contact area (10^{-12} – 10^{-10} cm²) and simple method of junction fabrication.

The good contact between crossed nanowires suggests that functional devices should be possible, and thus we have explored p–n junctions made from crossed p- and n-type nanowires. These junctions can be made reproducibly by sequential deposition of dilute solutions of n- and p-type nanowires with intermediate drying. Typical I – V data recorded on a crossed p–n junction (Fig. 2d) shows linear or nearly linear I – V for the individual n- and p-type nanowire components (green and blue curves), which indicates that they are ohmically contacted, and current rectification across the p–n junction (red curves); that is, little current flows in reverse bias, whereas there is a sharp current onset in forward bias. This diode-like behaviour is similar to bulk semiconductor p–n junctions, which form the basis for many critical electronic and

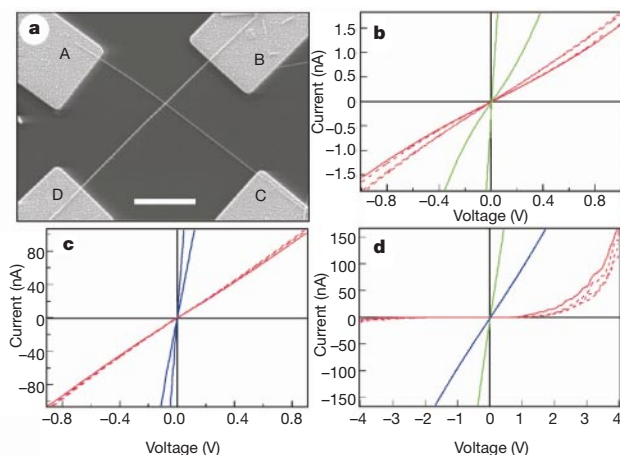


Figure 2 Crossed nanowire junctions and electrical properties. **a**, FE-SEM image of a typical crossed nanowire device with Ni/In/Au contact electrodes. The four arms are designated as A, B, C, D for simplicity of discussion. Scale bar, 2 μm. The diameters of the nanowires are 29 nm (AC) and 40 nm (BD); the diameters of the nanowires used to make devices were in the range of 20–75 nm. **b–d**, I – V behaviour of n–n, p–p and p–n junctions, respectively. The green and blue curves correspond respectively to the I – V behaviour of individual n- and p-type nanowires in the junctions. The red curves represent the I – V behaviour across the junctions. The current recorded for the p- and n-type nanowires in **d** has been divided by 10 for better viewing. Solid lines, transport behaviour across one pair of adjacent arms; dashed lines, transport behaviour across the other three pairs of adjacent arms. Data were recorded at room temperature.

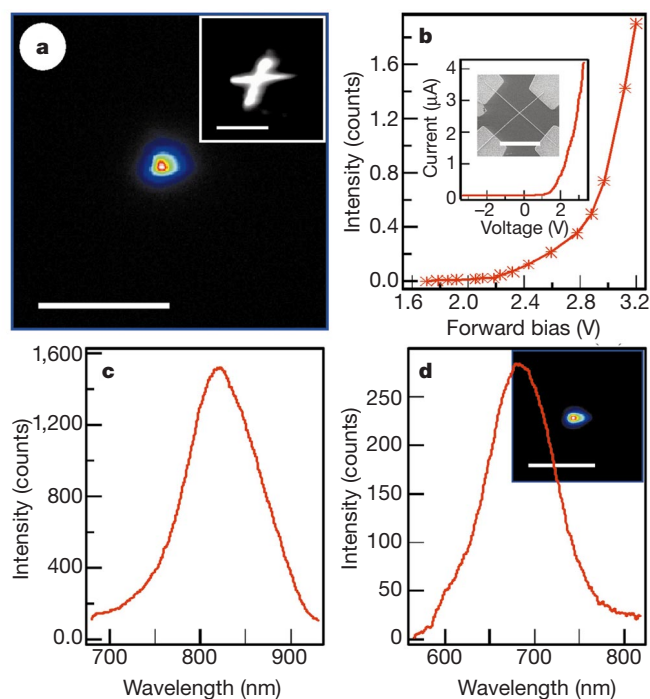


Figure 3 Optoelectrical characterization of nanowire p–n junctions. **a**, Electroluminescence (EL) image of the light emitted from a forward-biased nanowire p–n junction at 2.5 V. Inset, photoluminescence (PL) image of the junction. Scale bars, 5 μm. **b**, EL intensity versus voltage. Inset, I – V characteristics; inset in this inset, FE-SEM image of the junction itself. Scale bar, 5 μm. The n-type and p-type nanowires forming this junction have diameters of 65 and 68 nm, respectively. **c**, EL spectrum of the junction shown in **a**. The spectrum peaks at 820 nm. **d**, EL spectrum recorded from a second forward-biased crossed nanowire p–n junction. The EL maximum occurs at 680 nm. Inset, EL image showing that the EL originates from the junction region. Scale bar, 5 μm. The n-type and p-type nanowires forming this junction have diameters of 39 and 49 nm, respectively. The colours in **a** and the inset to **d** correspond to EL intensity, with black equal to zero counts and white equal to the highest counts.

optoelectronic devices. In a standard p–n junction, rectification arises from the potential barrier formed at the interface between p- and n-type materials¹². In the case of our nanowires, this picture may be modified, owing to a thin oxide at the interface (see Supplementary Information). When either type of junction is forward biased, the barrier is reduced and a relatively large current can flow through the junction; on the other hand, only a small current can flow in reverse bias as the barrier is further increased.

There are several reasons why the observed rectification can be attributed to the p–n junction formed between the crossed p- and n-type InP nanowires. First, the linear or nearly linear I – V behaviour of individual p- and n-type nanowires used to make the junction shows that ohmic contacts have been made between them and metal electrodes. This excludes rectification arising from metal–semiconductor Schottky diodes¹². Second, the I – V behaviour of the junction determined through every pair (AB, AD, BC, CD) of adjacent electrodes (red curves in Fig. 2d) exhibits a similar rectification effect and current level, which is also much smaller than the current level through the individual nanowires. These results demonstrate that the junction dominates the I – V behaviour. Third, four-terminal measurements in which current is passed through two adjacent electrodes (for example, AB), while the junction voltage drop is measured across two independent electrodes (for example, CD), exhibit similar I – V curves with only a slightly smaller voltage drop (0.1–0.2 V) compared to the two-terminal case (see Supplementary Information). Last, measurements made on 10 independent p–n junctions showed similar rectification in the I – V data; that is, significant current can only flow through p–n junctions when the p-type nanowire is positively biased.

The above data show unambiguously that we can rationally assemble nanoscale p–n junctions. In direct-bandgap semiconductors like InP, the p–n junction forms the basis for optoelectronics devices, including light-emitting diodes (LED) and lasers. To assess such functions in our nanoscale devices, we have studied the photoluminescence (PL) and electroluminescence (EL) from crossed

nanowire p–n junctions. Significantly, EL can be readily observed from these nanoscale p–n junctions in forward bias (Fig. 3a). A PL image of a crossed nanowire junction (inset) shows two crossed wire-like structures, and comparison of the EL and PL images shows that the position of the EL maximum corresponds to the crossing point in the PL image, thus demonstrating that the light originates from the nanowire p–n junction.

The I – V characteristics of the junction (Fig. 3b inset) shows clear rectification with a sharp current onset at ~ 1.5 V. The curve of EL intensity versus voltage shows (Fig. 3b) that significant light can be detected with our system at a voltage as low as 1.7 V. The EL intensity increases rapidly with bias voltage, and resembles the I – V behaviour. The EL spectrum (Fig. 3c) shows a maximum intensity around 820 nm, which is substantially blue-shifted relative to the bulk bandgap of InP (925 nm). EL results recorded from a p–n junction assembled using smaller diameter nanowires (Fig. 3d) showed a larger blue-shift, and thus suggest that these shifts may be due in part to quantum confinement¹³ of the excitons, although other factors also contribute (M.S. Gudiksen, J.W., X.D. and C.M.L., unpublished results). The quantum efficiency (electron to photon) of these initial devices is relatively low, $\sim 0.001\%$, which is not surprising as we have paid little attention to optimization. The efficiency is actually comparable to that ($\sim 0.002\%$) of early bulk InP LEDs¹⁴. We attribute the low quantum efficiency to non-radiative recombination via surface states, and believe that this deleterious process can be reduced by surface passivation^{15,16}.

Development of highly integrated devices based on nanowires will require techniques to align and assemble these building blocks into well-defined arrays. To demonstrate this next level of complexity, we have exploited electric fields to align and position individual nanowires into parallel and crossed arrays (Fig. 4a). A similar approach has been discussed¹⁷ elsewhere. The potential of this approach was first demonstrated by aligning many nanowires between parallel electrodes (Fig. 4b). FE-SEM images show that nearly all of the wires are aligned perpendicular to the parallel electrodes and along the direction of the electric field. Such electric-field assembly of nanowires between an array of electrodes (Fig. 4c) shows that individual wires can be positioned to bridge pairs of diametrically opposed electrodes and form a parallel array. In addition, by changing the field direction, the alignment can be done in a layer-by-layer fashion to produce crossed nanowire junctions (Fig. 4d). These data show that electric-field assembly represents a viable strategy for organizing individual nanowires with good directional and spatial control. We believe that highly integrated functional devices would be readily accessible using our nanowire building blocks in conjunction with this electric-field and/or other assembly techniques¹⁸.

Taken as a whole, the results presented here provide a rational approach for the bottom-up assembly of nanoscale electronic and optoelectronic devices. The demonstrated ability to assemble active devices in the absence of multi-billion-dollar fabrication lines is critically important to the field, and we believe that it augurs well for immediate and longer-term advances. We believe that the broad range of nanowire materials now available¹⁰ and the clearly defined ability to control their electronic properties will make possible nanoscale LEDs that cover the entire visible and near-infrared range (for example, GaN nanowires for blue colours¹⁹). Nanoscale light sources might be useful in creating new types of highly parallel sensors and for optical interconnects in nanoelectronics. We consider that the assembly of doped nanowire building blocks has great potential for creating many other types of electronic devices—and possibly even lasers. □

Methods

Nanowire synthesis and characterization

InP nanowires (NWs) were synthesized using LCG^{9,10}. The LCG target typically consisted of 94% (atomic ratio) InP, 5% Au as the catalyst, and 1% of Te or Zn as the doping element.

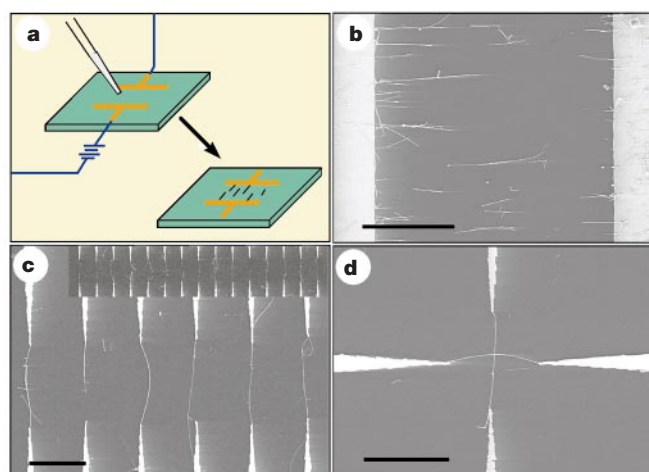


Figure 4 Parallel and orthogonal assembly of nanowires with electric fields. **a**, Schematic view of alignment by electric field. The electrodes (shown orange) are biased at 50–100 V after a drop of nanowire solution is deposited on the substrate (blue). **b**, Parallel array of nanowires aligned between two parallel electrodes. The nanowires were suspended in chlorobenzene and aligned using an applied bias of 100 V. **c**, Spatially positioned parallel array of nanowires obtained following electric-field assembly using a bias of 80 V. Inset, 15 pairs of parallel electrodes with individual nanowires bridging each diametrically opposed electrode pair. **d**, Crossed nanowire junction obtained using layer-by-layer alignment with the electric field applied in orthogonal directions in the two assembly steps. The applied bias in both steps was 80 V. Scale bars in **b–d**, 10 μm .

The furnace temperature (middle) was set at 800 °C during growth, and the target was placed at the upstream end rather than middle of the furnace. A pulsed (8-ns, 10-Hz) Nd-YAG laser (wavelength 1,064 nm) was used to vaporize the target. Typically, growth was performed for 10 min with NWs collected at the downstream, cool end of the furnace.

Electrical characterization

Transport measurement on individual NWs were carried out using published procedures¹¹. NWs were first dispersed in ethanol, and then deposited onto oxidized silicon substrates (600 nm oxide, 1–10 Ω cm resistivity), with the conductive silicon used as a back gate. Electrical contact to the NWs was defined using electron beam lithography (JEOL 6400). Ni/In/Au contact electrodes were thermally evaporated. Electrical transport measurements were made using a home-built system with <1 pA noise under computer control.

Junctions (n–n and p–p) were obtained by random deposition. We first deposited NWs onto oxidized silicon substrates using relatively high concentrations, determined the positions of crossed NWs, and then defined electrodes on all four arms of the cross by electron beam lithography. Ni/In/Au electrodes were used to make contact to the NWs.

p–n junctions were obtained by layer-by-layer deposition. First, a dilute solution of one type (for example, n-type) of NW was deposited on the substrate, and the positions of individual NWs were recorded. In a second step, a dilute solution of the other type (for example, p-type) of NW was deposited, and the positions of crossed n- and p-type NWs were recorded. Metal electrodes were then defined and transport behaviour was measured.

Optoelectrical characterization

EL was studied with a home-built micro-luminescence instrument²⁰. PL or scattered light (514 nm, Ar-ion laser) was used to locate the position of the junction. When the junction was located, the excitation laser was shut off, and then the junction was forward biased. EL images were taken with a liquid-nitrogen-cooled CCD camera, and EL spectra were obtained by dispersing EL with a 150 lines mm^{−1} grating in a 300-mm spectrometer.

Received 11 July; accepted 25 October 2000.

- Hu, J., Odom, T. W. & Lieber, C. M. Chemistry and physics in one dimension: synthesis and properties of nanowires and nanotubes. *Acc. Chem. Res.* **32**, 435–445 (1999).
- Dekker, C. Carbon nanotubes as molecular quantum wires. *Phys. Today* **52**(5), 22–28 (1999).
- Tans, S. J., Verschueren, R. M. & Dekker, C. Room temperature transistor based on a single carbon nanotube. *Nature* **393**, 49–52 (1998).
- Martel, R., Schmidt, T., Shea, H. R., Hertel, T. & Avouris, P. Single- and multi-wall carbon nanotube field effect transistors. *Appl. Phys. Lett.* **73**, 2447–2449 (1998).
- Tans, S. J. *et al.* Individual single-wall carbon nanotubes as quantum wires. *Nature* **386**, 474–477 (1997).
- Bockrath, M. *et al.* Single electron transport in ropes of carbon nanotubes. *Science* **275**, 1922–1925 (1997).
- Odom, T. W., Huang, J.-L., Kim, P. & Lieber, C. M. Atomic structure and electronic properties of single-walled carbon nanotubes. *Nature* **391**, 62–64 (1998).
- Wildoer, J. W. G., Venema, L. C., Rinzler, A. G., Smalley, R. E. & Dekker, C. Electronic structure of atomically resolved carbon nanotubes. *Nature* **391**, 59–62 (1998).
- Morales, A. M. & Lieber, C. M. A laser ablation method for the synthesis of crystalline semiconductor nanowires. *Science* **279**, 208–211 (1998).
- Duan, X. & Lieber, C. M. General synthesis of compound semiconductor nanowires. *Adv. Mater.* **12**, 298–302 (2000).
- Cui, Y., Duan, X., Hu, J. & Lieber, C. M. Doping and electrical transport in silicon nanowires. *J. Phys. Chem. B* **104**, 5213–5216 (2000).
- Sze, S. M. *Physics of Semiconductor Devices* (Wiley, New York, 1981).
- Alivisatos, A. P. Semiconductor clusters, nanocrystal, and quantum dots. *Science* **271**, 933–937 (1996).
- Bolm, G. M. & Woodall, J. M. Efficient electroluminescence from InP diodes grown by liquid-phase epitaxy. *Appl. Phys. Lett.* **17**, 373–376 (1970).
- Bessolov, V. N. & Lebedev, M. V. Chalco-genide passivation of III–V semiconductor surfaces. *Semiconductors* **32**, 1141–1156 (1998).
- Micic, O. I., Sprague, J., Lu, Z. & Nozik, A. J. Highly efficient band edge emission from InP quantum dots. *Appl. Phys. Lett.* **68**, 3150–3152 (1996).
- Smith, P. A. Electric-field assisted assembly and alignment of metallic nanowires. *Appl. Phys. Lett.* **77**, 1399–1401 (2000).
- Xia, Y. N., Rogers, J. A., Paul, K. E. & Whitesides, G. M. Unconventional methods for fabricating and patterning nanostructures. *Chem. Rev.* **99**, 1823–1848 (1999).
- Duan, X. & Lieber, C. M. Laser assisted catalytic growth of single crystal GaN nanowires. *J. Am. Chem. Soc.* **122**, 188–189 (2000).
- Duan, X., Wang, J. & Lieber, C. M. Synthesis and optical properties of GaAs nanowires. *Appl. Phys. Lett.* **76**, 1116–1168 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank H. Park, M.S. Gudiksen, J.-L. Huang, K. Kim, T. Oosterkamp & S.-I. Yang for discussions. This work was supported by the US Office of Naval Research, Defense Advanced Projects Research Agency, and the National Science Foundation.

Correspondence and requests for materials should be addressed to C.M.L. (e-mail: cml@cmliris.harvard.edu).

Computational design of direct-bandgap semiconductors that lattice-match silicon

Peihong Zhang*, Vincent H. Crespi*, Eric Chang†, Steven G. Louie† & Marvin L. Cohen†

* Department of Physics, The Pennsylvania State University, 104 Davey Lab, University Park, Pennsylvania 16802-6300, USA

† Department of Physics, University of California at Berkeley, Berkeley, California 94720 and Materials Sciences Division, Lawrence Berkeley Laboratory, Berkeley, California 94720, USA

Crystalline silicon is an indirect-bandgap semiconductor, making it an inefficient emitter of light. The successful integration of silicon-based electronics with optical components will therefore require optically active (for example, direct-bandgap) materials that can be grown on silicon with high-quality interfaces. For well ordered materials, this effectively translates into the requirement that such materials lattice-match silicon: lattice mismatch generally causes cracks and poor interface properties once the mismatched overlayer exceeds a very thin critical thickness. But no direct-bandgap semiconductor has yet been produced that can lattice-match silicon, and previously suggested structures¹ pose formidable challenges for synthesis. Much recent work has therefore focused on introducing compliant transition layers between the mismatched components^{2–4}. Here we propose a more direct solution to integrating silicon electronics with optical components. We have computationally designed two hypothetical direct-bandgap semiconductor alloys, the synthesis of which should be possible through the deposition of specific group-IV precursor molecules^{5,6} and which lattice-match silicon to 0.5–1% along lattice planes with low Miller indices. The calculated bandgaps (and hence the frequency of emitted light) lie in the window of minimal absorption in current optical fibres.

Bulk silicon is an indirect-bandgap semiconductor, that is the lowest-energy transition from valence to conduction bands involves a change in crystal momentum. Such materials are typically not suitable for optical applications since nonradiative events dominate the interband transitions. However, silicon is nearly surrounded by direct-bandgap elemental (unary) and compound semiconductors in the periodic table. Moving downwards from Si, the unary group-IV materials acquire larger cores containing *d* electrons; these states affect the conduction band states through orthogonality requirements and changes in the overall volume. On moving from silicon to germanium to tin, the $\Gamma_{2'}$ conduction band (as labelled in silicon) at *k* = 0 drops in energy until, in grey tin, the material acquires a direct (and vanishing) bandgap at *k* = 0. Moving across the periodic table, GaAs, a prototypical direct-bandgap semiconductor, differs from Si or Ge in that the constituents have different electronegativities. The resulting antisymmetric component in the crystal potential flattens the bands and opens the bandgap. In addition, the Γ_6 point of the conduction band drops relative to the other points in the band until it becomes the bottom of the conduction band in most of the well-known direct-bandgap group III–V and II–VI semiconductors.

These well-known results suggest that perhaps one can combine these mechanisms: larger cores and partial ionicity, within alloys containing only group-IV elements wherein the difference in rows of the periodic table provides the contrast in electronegativity. Such a system is likely to require the presence of tin to obtain a direct bandgap^{7–10}; the combination of tin (which is substantially larger than silicon) with carbon (which is substantially smaller than silicon) not only affords the greatest contrast in electronegativity

(which is needed to open the zero-bandgap of elemental Sn), but also opens the possibility of tuning the stoichiometry to obtain a lattice match to silicon. Also, the counteracting effects of the larger Sn atoms might compensate for the strain introduced by carbon within a Si/Ge lattice and therefore enhance the solubility of carbon within a silicon and germanium matrix.

To be practical, such an alloy must be accessible to a well-defined synthetic technique. One interesting previous study in computational design¹ has proposed a direct-bandgap lattice-match material. Unfortunately, this material requires the sequential deposition of eight distinct non-interdiffusing monolayers of As/Zn/As/Si/As/Zn/P/Si; this is a formidable synthetic challenge. At first sight, the current proposal for group-IV alloys seems also to present a forbidding synthetic challenge, because tetrahedrally coordinated alloys built from constituents of very different atomic sizes are unlikely to be thermodynamically stable. However, recent synthetic advances⁶ allow the production of semiconductor alloys with well-defined (and thermodynamically inaccessible) 1:4 C:Si and C:Ge stoichiometries through the use of precursor molecules which build in the required interatomic bonding, for example, $\text{CSi}_4\text{H}_{12}$. The terminal groups (such as H) are eliminated to produce the pure C/Si (or C/Ge) ordered alloy. We can thus produce bulk amounts of metastable alloys of controlled stoichiometry which are not accessible by standard near-equilibrium synthesis methods.

Of particular relevance to the current study are several additional synthetic developments. First, the Sn–D bond has recently been shown to be stable for long periods (at 0 °C) in molecules which incorporate Sn–D₃ moieties; (the Sn–H₃ moiety, in contrast, is known to be unstable). Second, techniques have been developed to

synthesize tetrahedral precursors where a central carbon atom is attached to four tin atoms¹¹. Finally, the synthetic yields of $\text{CSi}_4\text{H}_{12}$ and $\text{CGe}_4\text{H}_{12}$, for example, have improved significantly since their discovery, to the range where multi-gram quantities can be regularly prepared^{12,13}. Keeping these important advances in mind, we focus on unusual group-IV alloys which can be built from $(\text{Si}_4\text{Ge}_4\text{Sn})_4\text{X}_{12}$ precursors. In particular, we demonstrate that two alloys, CSi_2Sn_2 and CGe_3Sn , have low-energy allotropes which lattice-match silicon to better than ~0.5–1% and have direct bandgaps. We use the *ab initio* pseudopotential method¹⁴ in the local density approximation (LDA) to calculate both the structural and electronic properties of the proposed group-IV compounds. To obtain the energy gaps more precisely, we also calculate the quasi-particle energy gap using the GW approximation¹⁵.

Figure 1 shows the proposed molecular precursors for CSi_2Sn_2 and CGe_3Sn , namely, $\text{C}(\text{SiX}_3)_2(\text{SnX}_3)_2$ and $\text{C}(\text{GeX}_3)_3(\text{SnX}_3)$ (where X is a terminal group), and the corresponding ordered crystalline phases. These particular crystal structures are the lowest-energy (and highest-symmetry) from among several allotropes with small unit cells built from these precursor molecules. The structures shown are fully relaxed locally and all phonon modes are stable, indicating local structural stability. These $\text{C}(\text{AX}_3)_{4-x}(\text{BX}_3)_x$ molecular precursors maintain the desired tetrahedral bonding. The $\text{CA}_{4-x}\text{B}_x$ molecular cores also provide the highest possible carbon concentrations without creating energetically unfavourable nearest-neighbour or second-nearest-neighbour C–C pairs¹⁶. In addition, the absence of strong C–X (such as C–H) bonds allows a low growth

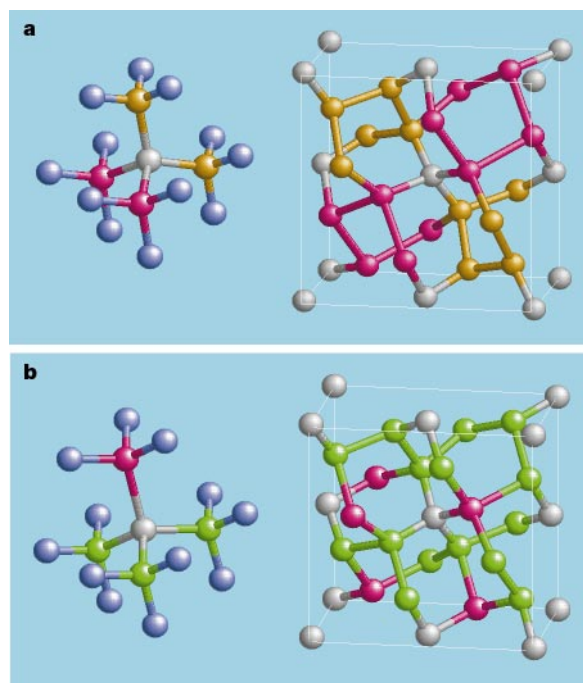


Figure 1 Proposed molecular precursors and the corresponding relaxed crystalline phases of the new group-IV compounds. The crystalline structures shown here are the highest in symmetry among several possible allotropes, and thus are the most plausible structures assumed when depositing on a silicon substrate as discussed in the text. Silicon is yellow, germanium is green, tin is magenta, carbon is grey, and the termination group is light blue. **a**, The molecular precursor $\text{C}(\text{SiX}_3)_2(\text{SnX}_3)_2$ and the corresponding crystal structure CSi_2Sn_2 . **b**, The precursor $\text{C}(\text{GeX}_3)_3(\text{SnX}_3)$ and the crystal CGe_3Sn . CSi_2Sn_2 has a body-centred tetragonal (b.c.t.) lattice. The three lattice constants are: $a = b = 8.43 \text{ \AA}$, $c = 5.46 \text{ \AA}$. The lattice of CGe_3Sn is slightly distorted from b.c.t., with lattice constants $a = 8.50 \text{ \AA}$, $b = 8.33 \text{ \AA}$, and $c = 5.41 \text{ \AA}$. The angle between a and b is 89.6° .

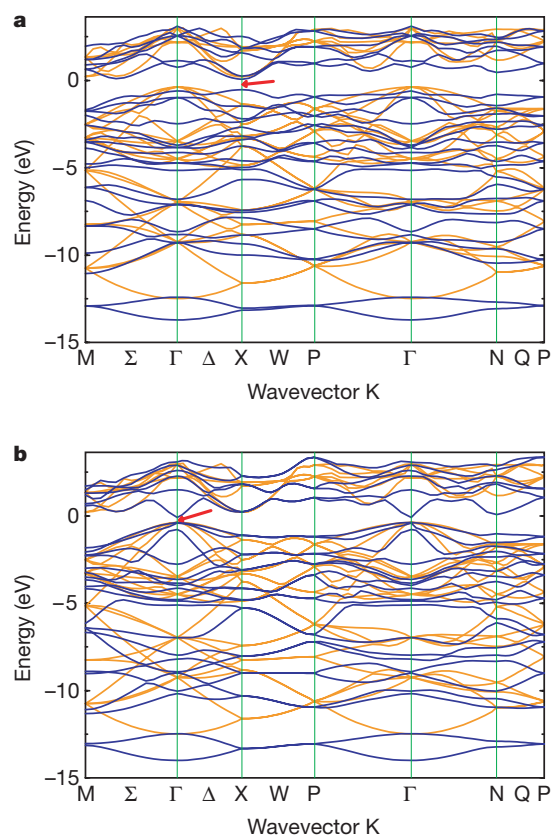


Figure 2 Band structures of novel group-IV alloys compared to that of silicon. The figure shows band structures (in blue) of **a**, CSi_2Sn_2 and **b**, CGe_3Sn . The ionicity opens a bandgap deep in the valence band¹⁹. Symmetry breaking in CSi_2Sn_2 removes the degeneracy of the X point, pushing one band up significantly so that it becomes the conduction-band maximum and produces a direct band gap at X (red arrow). In the case of CGe_3Sn , the conduction-band minimum occurs at Γ which produces a direct gap at Γ (red arrow). The band structure of silicon (in orange) calculated in the same unit cell is also plotted for comparison.

temperature which is crucial for synthesizing metastable materials, particularly with tetrahedrally coordinated Sn.

The CSi_2Sn_2 material has a body-centred tetragonal (b.c.t.) lattice with 10 atoms per unit cell. This relatively high-symmetry structure can minimize the anisotropic strains associated with deposition on a low-index silicon surface. The fully relaxed density functional results reveal that the volume per atom in CSi_2Sn_2 is only 0.6% less than that of bulk silicon; (we compare the theoretical LDA volumes to minimize systematic errors). This is only a slight deviation from Vegard's law, which predicts that stoichiometries near $\text{Si}_{2.22}\text{Sn}_{1.78}\text{C}_1$ would lattice-match silicon. The b.c.t. structure of CSi_2Sn_2 provides several lattice planes which can be well-matched to low-index planes of bulk silicon. For example, the (110) and (111) planes of silicon match to CSi_2Sn_2 better than 0.5% in all directions. The 3:1 ratio of noncarbon atomic constituents in CGe_3Sn yields a slightly lower-symmetry structure which is 1.7% lower in volume than bulk silicon. Owing to the lower crystal symmetry, the lattice is slightly distorted from b.c.t. and the best match to silicon is somewhat worse, but still less than 1% in linear dimension (with a combination of compression and dilation along different axes). Mismatches of this order should yield a critical thickness of several hundred or thousand ångströms (see ref. 17 for example). In addition, a synthesis based on these molecular precursors could allow the addition of a small concentration of a second precursor that helps to tune the lattice match precisely.

Figure 2 compares the pseudopotential density functional band-structures of CSi_2Sn_2 and CGe_3Sn to the corresponding Si band-structure within the same (ten-atom) unit cell. (The specific points X and so on, mentioned below, refer to this ten-atom b.c.t. cell and not the standard two-atom silicon Brillouin zone). CSi_2Sn_2 has a direct bandgap in an unusual position, the X point. The conduction-band minimum at the X point is rather close to the folded X point in Si. Symmetry breaking due to the alloying splits a degeneracy at the X point. In the uppermost valence band, this splitting raises the higher-energy state substantially until it passes the Γ -point state and thereby becomes the new valence-band maximum. This change in the valence-band maximum compared to Si is facilitated by the smaller bandwidths in this more ionic semiconductor alloy. Meanwhile, the conduction band at Γ follows a well-known pattern; the introduction of Sn pulls down the lowest conduction band at Γ and produces a local (but in this case not global) conduction-band minimum at Γ . To understand this somewhat abnormal band structure, that is, a direct energy gap at a point other than Γ , we analyse the charge density at the top valence band for the X and Γ points. At the Γ point, electronic charge is preferentially concentrated around Si–Si bonds. In contrast, at the X point the charge density is higher around the Sn–Sn bonds. Measuring the bond charge contained within fixed-width cylinders, we obtain $(Q_\Gamma(\text{Si–Si})/Q_\Gamma(\text{Sn–Sn})) : (Q_X(\text{Si–Si})/Q_X(\text{Sn–Sn})) \approx 4 : 1$ (We define the bond charge as the total charge enclosed by a cylinder connecting the two atoms, denoted by the elemental symbol in italics. Although the absolute value of the bond charge scales with the volume of the cylinder, the ratio of the bond charges is insensitive to the cylinder diameter.) Because the valence atomic-orbital energies of Sn are higher than those of Si, the greater Sn character of the X point accounts for the unusual valence-band structure of CSi_2Sn_2 . The fundamental LDA bandgap of CSi_2Sn_2 at the X point is 0.59 eV. The GW correction increases the bandgap to 0.9 eV.

The mechanisms which produce a direct bandgap in CGe_3Sn have a more familiar interpretation. Owing to the increased fraction of the lower-row elements Ge and Sn, the drop in the conduction band at Γ is more pronounced than in CSi_2Sn_2 . This Γ -point state becomes the new conduction-band minimum, directly above the valence-band maximum, which remains at Γ . The LDA bandgap at Γ is 0.30 eV, increasing to 0.71 eV under the GW correction. The fundamental band gap at Γ is very sensitive to external pressure,

dropping by 0.1 volts for a fractional change in volume of $\Delta V/V = 0.01$. This high sensitivity arises from the volume variation in the local potential energy of the conduction band¹⁸ and provides a potentially important mechanism for tuning the size of the direct bandgap.

For consistency, the band structures are calculated at the experimental volume per atom of the posited silicon substrate. Because the volume differences between these materials and bulk silicon are very small, the band structures are very similar to those calculated at their equilibrium volumes. CSi_2Sn_2 shows no significant changes in band structure under such a volume change. The high volume sensitivity of CGe_3Sn means that the small expansion to match the experimental volume per atom of silicon actually drives the band-gap to be direct. (At the relaxed bulk CGe_3Sn volume the bandgap is just slightly indirect.) For CGe_3Sn the optical matrix elements at the direct transition are comparable to those in GaAs. For CSi_2Sn_2 , the optical matrix elements are somewhat smaller but the transition is still clearly dipole-allowed.

We have shown how materials design can yield several candidate materials which simultaneously possess three desirable materials properties: a very close lattice match to silicon, a direct band-gap in the 0.7–1.0 V range, and accessibility to well-defined synthetic strategies that can potentially produce bulk quantities of material. □

Received 28 September; accepted 2 November 2000.

- Wang, T., Moll, N., Cho, K. & Joannopoulos, J. D. Deliberately designed materials for optoelectronics applications. *Phys. Rev. Lett.* **82**, 3304–3306 (1999).
- Ejcek, F. E., Lo, Y. H., Subramanian, S., Hou, H. Q. & Hammons, B. E. Lattice engineered compliant substrate for defect-free heteroepitaxial growth. *Appl. Phys. Lett.* **70**, 1685–1687 (1997).
- Lo, Y. H. New approach to grow pseudomorphic structures over the critical thickness. *Appl. Phys. Lett.* **59**, 2311–2313 (1991).
- Powell, A. R., Iyer, S. S. & LeGoues, F. K. New approach to the growth of low dislocation relaxed SiGe material. *Appl. Phys. Lett.* **64**, 1856–1858 (1994).
- Kouvetakis, J., Nesting, D. & Smith, D. J. Synthesis and atomic and electronic structure of new Si–Ge–C alloys and compounds. *Chem. Mater.* **10**, 2935–2949 (1998).
- Kouvetakis, J., Chandrasekhar, D. & Smith, D. J. Growth and characterization of thin $\text{Si}_{80}\text{C}_{20}$ films based upon Si_3C_2 building blocks. *Appl. Phys. Lett.* **72**, 930–932 (1998).
- Jenkins, D. W. & Dow, J. D. Electronic properties of metastable $\text{Ge}_x\text{Sn}_{1-x}$ alloys. *Phys. Rev. B* **36**, 7994–8000 (1987).
- Mäder, K. A., Baldereschi, A. & von Känel, H. Band structure and instability of $\text{Ge}_{1-x}\text{Sn}_x$ alloys. *Solid State Commun.* **69**, 1123–1126 (1989).
- He, G. & Atwater, H. A. Interband transitions in $\text{Sn}_x\text{Ge}_{1-x}$ alloys. *Phys. Rev. Lett.* **79**, 1937–1940 (1997).
- Min, K. S. & Atwater, H. A. Ultrathin pseudomorphic Sn/Si and $\text{Sn}_x\text{Si}_{1-x}$ /Si heterostructures. *Appl. Phys. Lett.* **72**, 1884–1886 (1998).
- Klinkhammer, K. W., Kühner, S., Regelman, B. & Weidlein, J. Die Kristall- und Molekülstruktur von Tetrakis(trimethylstannyl)methan. *Organomet. Chem.* **496**, 241–243 (1995).
- Schmidbauer, H. & Zech, J. An improved synthetic pathway to tetrasilyl methane and the synthesis of disilylpropane. *Eur. J. Solid State Inorg. Chem.* **29**, 5–21 (1992).
- Kouvetakis, J. *et al.* Novel methods for CVD of Ge_xC and $(\text{Ge}_x\text{C})_y\text{Si}_z$ diamond-like semiconductor heterostructures: Synthetic pathways and structures of trigermyl- $(\text{GeH}_3)_3\text{CH}$ and tetragermyl- $(\text{GeH}_3)_4\text{C}$ methanes. *J. Am. Chem. Soc.* **120**, 6738–6744 (1998).
- Ihm, J., Zunger, A. & Cohen, M. L. Momentum-space formalism for the total energy of solids. *J. Phys. C* **12**, 4409–4423 (1979).
- Hybertsen, M. S. & Louie, S. G. Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies. *Phys. Rev. B* **34**, 5390–5413 (1986).
- Rücher, H., Metthessel, M., Bugiel, E. & Osten, H. J. Strain-stabilized highly concentrated pseudomorphic $\text{Si}_{1-x}\text{C}_x$ layers in Si. *Phys. Rev. Lett.* **72**, 3578–3581 (1994).
- Herman, M. A. & Sitter, H. *Molecular Beam Epitaxy: Fundamentals and Current Status* (Springer, Berlin/Heidelberg, 1989).
- Corkill, J. L. & Cohen, M. L. Band gaps in some group-IV materials: A theoretical analysis. *Phys. Rev. B* **47**, 10304–10309 (1993).
- Cohen, M. L. & Chelikowsky, J. R. *Electronic Structure and Optical Properties of Semiconductors* (Springer, Berlin/Heidelberg, 1988).

Acknowledgements

V.H.C. thanks T. Mallouk and J. Kouvetakis for useful discussions and J. Kouvetakis for information on the stability of Sn-D_3 moiety. V.H.C. acknowledges support from the Packard Foundation and from the National Science Foundation, Division of Materials Research. V.H.C. also acknowledges the National Partnership for Advanced Computational Infrastructure and the Pittsburgh Supercomputing Center for computational support. M.L.C. and S.G.L. acknowledge support from the National Science Foundation, Division of Materials Research and from the Office of Energy Research, Office of Basic Energy Sciences, Materials Sciences Division of the US Department of Energy.

Correspondence and requests for materials should be addressed to V.H.C. (e-mail: crespi@phys.psu.edu).

Evidence from gabbro of the Troodos ophiolite for lateral magma transport along a slow-spreading mid-ocean ridge

Meir Abelson^{*†}, Gidon Baer^{*} & Amotz Agnon[†]

^{*} Geological Survey of Israel, 30 Malkhey Yisrael Street, Jerusalem 95501, Israel

[†] Institute of Earth Sciences, Hebrew University, Jerusalem 91904, Israel

The lateral flow of magma and ductile deformation of the lower crust along oceanic spreading axes has been thought to play a significant role in suppressing both mid-ocean ridge segmentation^{1,2} and variations in crustal thickness^{3,4}. Direct investigation of such flow patterns is hampered by the kilometres of water that cover the oceanic crust, but such studies can be made on ophiolites⁵ (fragments of oceanic crust accreted to a continent). In the Oman ophiolite, small-scale radial patterns of flow have

been mapped along what is thought to be the relict of a fast-spreading mid-ocean ridge⁵. Here we present evidence for broad-scale along-axis flow that has been frozen into the gabbro of the Troodos ophiolite in Cyprus (thought to be representative of a slow-spreading ridge axis). The gabbro suite of Troodos spans nearly 20 km of a segment of a fossil spreading axis, near a ridge-transform intersection^{6,7}. We mapped the pattern of magma flow by analysing the rocks' magnetic fabric at 20 sites widely distributed in the gabbro suite, and by examining the petrographic fabric at 9 sites. We infer an along-axis magma flow for much of the gabbro suite, which indicates that redistribution of melt occurred towards the segment edge in a large depth range of the oceanic crust. Our results support the magma plumbing structure that has been inferred indirectly from a seismic tomography experiment on the slow-spreading Mid-Atlantic Ridge⁸.

In the past decade, the understanding of plumbing in magmatic systems in fast-spreading ridges (for example, the East Pacific Rise, EPR) has been improved significantly by geophysical observations of the present active EPR (see, for example, ref. 9 for a review) and geological studies of the Oman ophiolite⁵. On the other hand, the melt course and magmatic structure in slow-spreading axes (for example, the Mid-Atlantic Ridge, MAR, and the southwest Indian

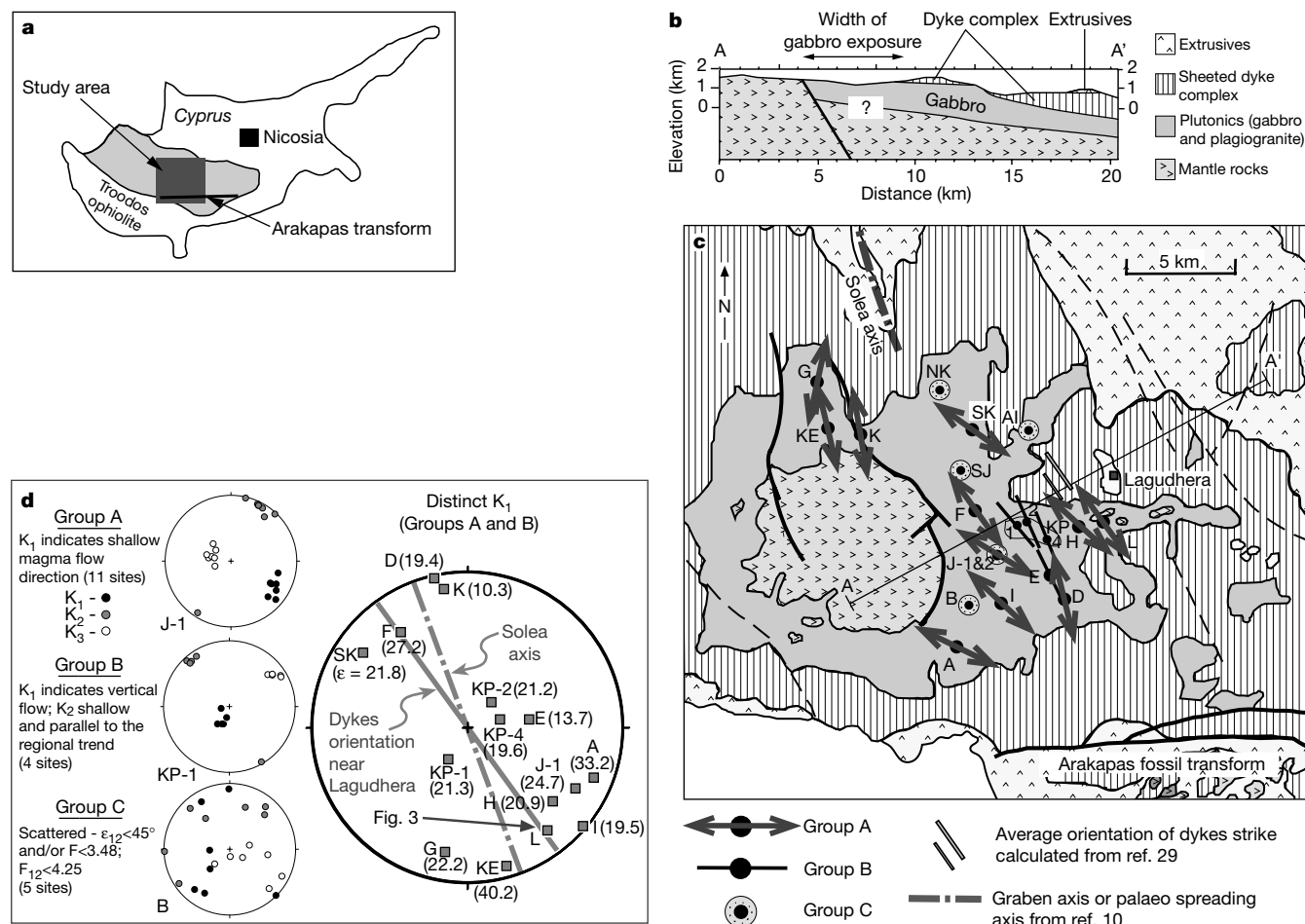


Figure 1 Location maps and results of AMS. **a**, Location map; **b**, cross-section at Troodos, Cyprus. **c**, Location of 20 sampling sites in the gabbro suite on a geological map²⁷ of the area around the Solea graben. **d**, Site mean directions of K_1 axes plotted on a lower-hemisphere equal-area stereographic projection. Mean directions were calculated using Pmag software²⁸, which applies Hext linear perturbation analysis, and further gives (1) the semi-major axis (ϵ_1) of the confidence ellipse for K_1 towards K_2 (numbers in brackets next to each site), and (2) F-statistics for testing isotropy (F) and distinction of K_1 (F_{12}) (F_{12} is small for high statistical variance and an oblate AMS ellipsoid). Flow directions could not be inferred for 5 sites (group C), where $\epsilon > 45^\circ$ (K_1 s are too scattered) and/or

$F < 3.48$; $F_{12} < 4.25$ for this group (ref. 28). In addition to these 5 sites, 41 samples were excluded from other sites (29 from site L alone, see Fig. 3) due to weak F and F_{12} . Fifteen sites show clear preferred orientations of AMS axes (groups A and B). We note parallelism between shallow K_1 (group A, thick arrows on the map), orientation of dykes (taken from ref. 29) and the Solea axis. The degree of anisotropy, P' , and the shape factor (oblate, triaxial or prolate) were also calculated by the Jelinek method³⁰. An overwhelming majority of the measured samples have a prominent degree of anisotropy ($1.02 < P' < 1.2$ for all sites, $P' = 1.065 \pm 0.025$ in the non-layered gabbro), and a triaxial shape of the AMS ellipsoid.

ridge) have been poorly constrained. A recent seismic tomography study of the slow-spreading MAR⁸ has revealed a vertical, cylinder-like seismic velocity anomaly beneath a segment mid-point, which was interpreted as a hot region or the presence of partial melt, presumably a feeding conduit. Similar anomalies were spread at shallow depths along the spreading axis and may reflect ponding of melt⁸. But the detailed pattern and level of magmatic flow (gabbro versus sheeted dykes) remain to be resolved.

The Solea graben of the Troodos ophiolite has been considered as an exposed example of a slow^{6,10,11}-to-intermediate¹¹ spreading axis.

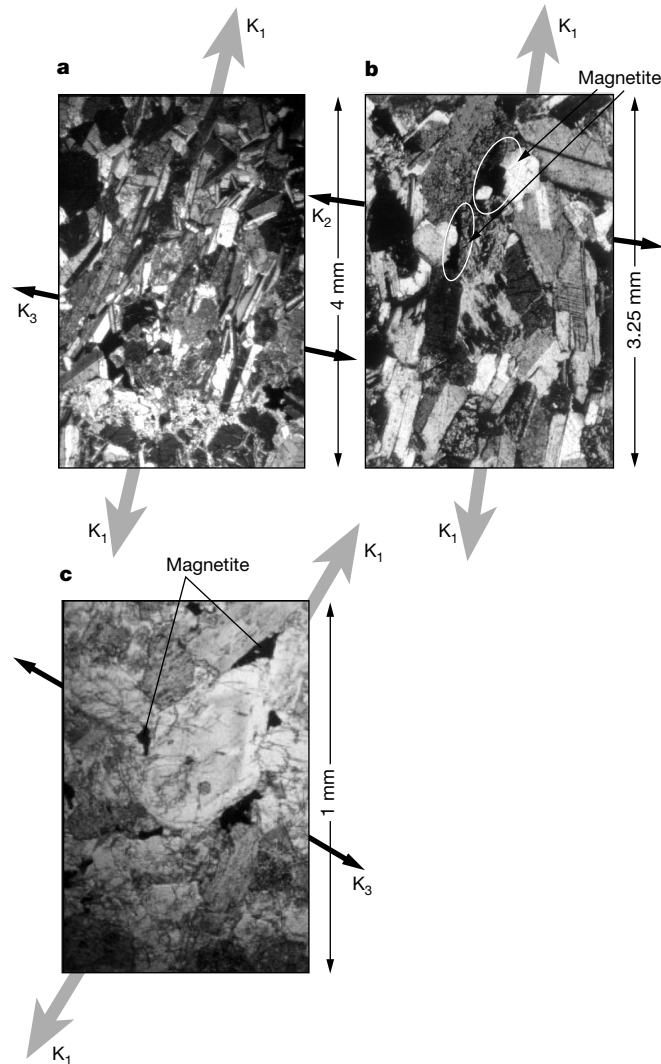


Figure 2 The relation between magnetic fabric and photographic fabric. **a, b.** Photographs of polarized thin sections from a non-layered gabbro (site KP-2), cut to show K_1 – K_3 (**a**) and K_1 – K_2 (**b**) principal susceptibility axes. We note that K_1 and mineral elongation are parallel in both sections. We prepared 33 orthogonal thin sections paralleling the AMS principal planes from 9 selected sites. The relation between magnetic fabric and petrographic fabric displayed here are observed in 6 of the 9 sites (A, D, J-1, K, KP-2, L-3 of site L). In two sites (J-2 and SJ, group C), no flow fabric is observed, as also indicated by their respective scattered K_1 . These two sites are also exceptional due to the observed brittle deformation. The rest of the sites are devoid of solid-state deformation, as typically observed in the Troodos gabbro^{12,14}. Two sites (E and L-2 of site L) display pronounced flow fabric in the K_1 – K_3 and K_2 – K_3 planes with no preferred orientation in the K_1 – K_2 plane (K_3 is normal to microscopic foliation). **c.** A non-polarized thin section showing magnetite that crystallized interstitially within a pre-existing preferentially oriented rock fabric. This may suggest that the origin of AMS is the distribution anisotropy¹⁷ of magnetite grains, in agreement with the measured high susceptibilities, and as previously inferred for abyssal gabbros¹⁸.

The gabbro suite at the Troodos ophiolite is exposed between the extinct spreading axis of the Solea graben and the east–west-trending fossil oceanic transform (the Arakapas transform; Fig. 1). The extinct spreading axis in the Solea region was identified by the orientation of dykes and by the opposing tilt of dykes and faulted blocks^{7,10}. The distance between the northernmost gabbroic

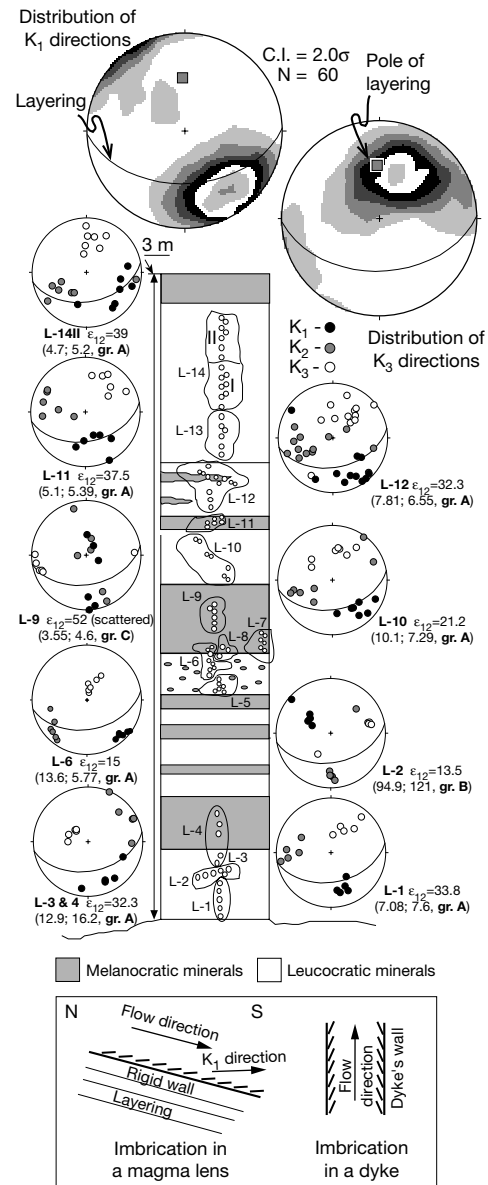


Figure 3 Directions of AMS principal axes across the section of the layered gabbro in site L (Lagudhera). This gabbro displays cumulate features¹³, implying layering formation in a magma pond with no late transposition and deformation. 77% of the rock samples have shallow K_1 parallel to the northwest–southeast regional trend. The thin lines encircling the drill-holes define rock samples that were drilled in the laboratory at the Geological Survey of Israel. AMS measurements (97) were conducted on these 14 rock samples. Thirty-seven measurements of 5 rock samples (L-5, L-7, L-8, L-13 and L-14) were excluded due to weak F and F_{12} (Fig. 1) (2 excluded measurements are outliers in L-6). The remaining measurements (60 of 9 rock samples) are displayed with the length of the semi-major axis of the confidence ellipse of K_1 (ϵ_{12}), F and F_{12} (numbers in brackets next to stereoplots). K_1 and K_3 orientations are contoured by the Kamb method (top). We note that the K_1 s are mostly imbricated to the mineral layering, and K_3 deviates slightly from the normal to layering. This configuration indicates imbrication of the magnetic foliation of the fossil wall of the magma chamber (K_2 approximately parallels the layering strike). Similar AMS imbrication relations are commonly used to infer the flow sense in dykes²² (see inset). Accordingly, this imbrication suggests a southward flow in this site.

exposures, where they truncate the Solea axis, and the Arakapas fossil transform is ~20 km. This allows the detection of flow patterns over a region that is one-quarter to one-third of the length of a typical slow-spreading segment, close to a fossil ridge–transform intersection^{6,7}. Furthermore, the gabbro is exposed across various depth levels, from the uppermost part immediately below the sheeted dykes down to deeper levels close to the Moho (Fig. 1).

Most of the exposed gabbro in Troodos is non-layered¹². Layered gabbro is generally found at the lower level of the suite, but also occasionally at higher levels¹², even right beneath the dyke complex¹³ (for example, site L; Fig. 1). The Troodos gabbro is characterized by the lack of evidence for solid-state deformation^{12,13}. Field evidence and geochemistry further indicate that the gabbro suite was formed by cyclic events generating multiple magma chambers at various levels within the crust and in the uppermost mantle¹⁴.

In the past two decades, anisotropy of magnetic susceptibility (AMS) has been used extensively to detect magmatic flow directions (see ref. 15 for a review). The second-rank susceptibility tensor, K , may be visualized as an ellipsoid with three principal axes ($K_1 > K_2 > K_3$). The magnitude and orientation of this ellipsoid presumably reflect alignment of magnetic minerals due to shear during magma flow. The AMS technique efficiently assists the deduction of flow fabric in many sampling sites where the petrographic fabric is weakly present and difficult to map¹⁶. Here we report AMS results that reconstruct magma flow directions in the axial magma chambers frozen within the gabbro.

We performed AMS measurements on 236 samples using an SI-2 (Sapphire Instruments) instrument to determine low-field magnetic susceptibility and anisotropy. We used 20 sampling sites widely distributed within the gabbro suite to sample layered gabbro (sites G, K, KE and L) and macroscopically isotropic gabbro (all others) (Fig. 1). Of these 20 sites, 15 (groups A and B; Fig. 1) show a clear preferred orientation of the magnetic fabric. Of these 15 sites, 11 (group A) display shallow K_1 that is parallel to the regional northwest–southeast to north–south trend of the Solea graben axis, and to the orientation of dykes near the village of Lagudhera (Fig. 1). Whereas these 11 sites are widely distributed in the gabbroic suite, vertical K_1 is found in 4 sites (group B) concentrated in one region, 2–3-km wide (KP and E, Fig. 1). In the remaining 5 sites the AMS axes are randomly scattered (group C) (Fig. 1).

The AMS of most rock-forming minerals is coaxial to the petrographic fabric^{15,17,18}, as also found in abyssal gabbros (from the EPR¹⁸ and the Oman ophiolite¹⁹). The source of AMS in these gabbros is late crystallization of magnetite that conforms to the existing flow fabric (distribution anisotropy¹⁷), and K_1 usually

indicates flow direction. We have found further support for these relations in the Troodos ophiolite by comparing the petrographic fabric with the magnetic fabric: orthogonal thin sections cut so as to include AMS principal axes were prepared from 9 sites (Fig. 2). The clearest flow texture, expressed by preferred orientation of elongated minerals, is always seen on planes comprising K_1 and K_3 axes where K_1 parallels the elongation of the minerals (Fig. 2). Parallelism between mineral elongation and K_1 is also seen on K_1 – K_2 sections in most of the sites examined (6 sites out of 9, Fig. 2), further confirming K_1 as a flow indicator.

Accepting that K_1 is a reliable indicator of flow direction, the primarily shallow K_1 directions in the gabbro suggest that the magmatic flow within the axial palaeo magma chambers was predominantly lateral. This lateral flow is parallel to the extinct spreading axis (that is, parallel to the Solea axis and dykes near Lagudhera) along ~20 km of a segment end, north of the fossil Arakapas transform (Fig. 1).

In order to examine small-scale coherence and explore temporal variations in flow patterns, we took samples across a detailed columnar section of the layered gabbro near Lagudhera (site L; Fig. 3). Layering in such unfoliated, cumulate gabbro¹³ may represent a migrating solidification front in the periphery of a magma chamber²⁰. Therefore, an AMS profile through the section of this layered gabbro (Fig. 3) may record temporal variations of the magmatic flow. The results show that 77% of the samples show shallow K_1 sub-parallel to the northwest–southeast regional trend. Therefore, it seems that the flow pattern remains stable across the entire section, and is similar to the regional trend of the gabbro suite. Furthermore, K_1 axes are imbricated relative to the gabbro layering (Fig. 3 inset). If layering of the cumulate gabbro represents the migrating front of a rheological contrast between stronger and weaker layers²¹ on a floor of a fossil magma pond, the sense of imbrication suggests lateral flow southwards, towards the transform (Figs 1 and 3).

The southerly sense of flow in the gabbro agrees with southward trend of flow previously inferred in the sheeted dykes by AMS imbrication and other structural indicators^{22,23}. Flow from the mid-point of the segment to its edge is also in agreement with observations from modern spreading centres, where magmatism in the segment centre is more intense than at its edge^{8,9}.

The axis-parallel pattern of magma flow that we find in the Troodos gabbro differs from the radial pattern revealed in the gabbro of Oman ophiolite. Structural studies of the Oman ophiolite show diverging flow lines around a feeder region of a mantle diapir, also at the segment edge⁵. The absence of a radial flow pattern in Troodos, and hence the absence of a significant feeding feature, suggests that the origin of the magma could be more than 20 km north of the fossil oceanic transform. This is in agreement with the interpretation of MAR seismological data in ref. 8, where a feeder was inferred at the mid-point of a 90-km-long slow-spreading segment. Even so, we deduce some vertical magma flow in a zone ~2–3 km wide (sites KP and E, Fig. 1). This may represent (1) a local upwelling within a laterally propagating intrusion, or (2) a feeder zone (Fig. 4). Such a vertically flowing zone, surrounded by an axis-parallel flow field (Fig. 1), is not likely to be a major conduit, where radial divergence of flow lines would be expected (as found in Oman and subaerial volcanic edifices). We note that the anisotropy found in our results conforms with the expected axial plane, where the dominant direction is lateral and axis parallel.

Geophysical observations that indicate culmination of melt supply at the mid-point of a slow-spreading segment^{8,24} cannot completely resolve whether the melt redistributes along-axis at a particular crustal level or ascends upwardly from the mantle with varying fluxes along the entire segment. The direct evidence of along-axis melt migration reported here elucidates the course of melt within a slow-spreading axis at the crustal level: the melt is largely supplied to the segment mid-point, from where it migrates

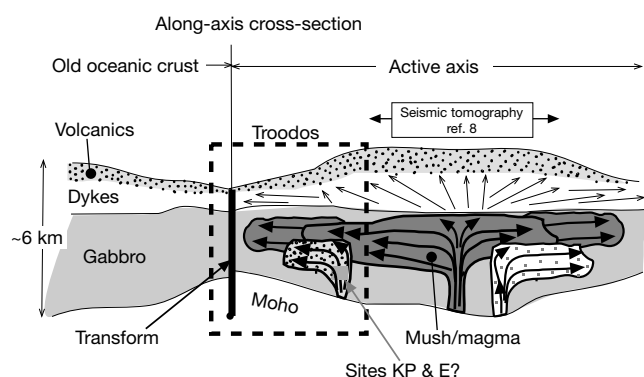


Figure 4 Cross-section along the palaeo spreading axis of Troodos of a simplified model. On the basis of field relations¹⁴ and the present AMS results, the melt in the gabbro layer is shown flowing laterally in multiple magmatic ponds. The sense of flow is from the segment centre towards the transform, as supported by the sense of the lateral flow found in the diabase dykes²³ and site L. Recent work⁸ underlines the along-axis variability within the central zone of a segment in a slow-spreading axis.

laterally along the segment axis to the segment end, within the dykes²³ and the gabbro layer (Fig. 4). As previously noted, along-axis melt redistribution has a significant role in controlling along-axis variations of crustal thickness^{4,25} and the topography of the ocean floor²⁶. Our results show that lateral flow in the gabbro layer, the most voluminous fraction of the oceanic crust (about two-thirds of the crust thickness), constitutes a considerable portion of the along-axis magma transport. □

Received 11 January; accepted 2 November 2000.

- Bell, R. E. & Buck, W. R. Crustal control of ridge segmentation inferred from observations of the Reykjanes Ridge. *Nature* **357**, 583–586 (1992).
- Abelson, M. & Agnon, A. Mechanics of oblique spreading and ridge segmentation. *Earth Planet. Sci. Lett.* **148**, 405–421 (1997).
- Lin, J. & Phipps-Morgan, J. The spreading rate dependence of 3-D mid-ocean gravity structure. *Geophys. Res. Lett.* **19**, 13–16 (1992).
- Wang, X. & Cochran, J. R. Along-axis gravity gradients at mid-ocean ridges: implications for mantle flow and axial morphology. *Geology* **23**, 29–32 (1995).
- Nicolas, A. & Boudier, F. Mapping oceanic ridge segments in Oman ophiolite. *J. Geophys. Res.* **100**, 6179–6197 (1995).
- Hurst, S. D., Karson, J. A. & Moores, E. M. in *Proc. Symp. "TROODOS 1987" on Ophiolites and Oceanic Crustal Analogues* (eds Malpas, J., Moores, E. M., Panayiotou, A. & Xenophontos, C.) 125–130 (Geological Survey Dept, Ministry of Agriculture and Natural Resources of Cyprus, Nicosia, 1990).
- MacLeod, C. J., Allerton, S., Gass, I. G. & Xenophontos, C. Structure of a fossil ridge-transform intersection in the Troodos ophiolite. *Nature* **348**, 717–720 (1990).
- Magde, L. S., Barclay, A. H., Toomey, D. R., Detrick, R. S. & Collins, J. A. Crustal magma plumbing within a segment of the Mid-Atlantic Ridge, 35°N. *Earth Planet. Sci. Lett.* **175**, 55–68 (2000).
- Solomon, S. C. & Toomey, D. R. The structure of mid-ocean ridges. *Annu. Rev. Earth Planet. Sci.* **20**, 329–364 (1992).
- Hurst, S. D., Moores, E. M. & Varga, R. J. Structural and geophysical expression of the Solea graben, Troodos Ophiolite, Cyprus. *Tectonics* **13**, 139–156 (1994).
- Varga, R. J., Gee, J. S., Bettison-Varga, L., Anderson, R. S. & Johnson, C. L. Early establishment of seafloor hydrothermal systems during structural extension: paleomagnetic evidence from the Troodos ophiolite, Cyprus. *Earth Planet. Sci. Lett.* **171**, 221–235 (1999).
- George, R. P. Structural petrology of the Olympus ultramafic complex in the Troodos ophiolite, Cyprus. *Geol. Soc. Am. Bull.* **89**, 845–865 (1978).
- Malpas, J., Xenophontos, C. & Cann, J. R. in *Third Int. Conf. on the Geology of the Eastern Mediterranean* (Nicosia, Cyprus, 1998).
- Malpas, J. in *Proc. Symp. "TROODOS 1987" on Ophiolites and Oceanic Crustal Analogues* (eds Malpas, J., Moores, E. M., Panayiotou, A. & Xenophontos, C.) 65–74 (Geological Survey Dept, Ministry of Agriculture and Natural Resources of Cyprus, Nicosia, 1990).
- Rochette, P., Jackson, M. & Aubourg, C. Rock magnetism and the interpretation of anisotropy of magnetic susceptibility. *Rev. Geophys.* **30**, 209–226 (1992).
- Ernst, R. E. & Baragar, W. R. A. Evidence from magnetic fabric for the flow pattern of magma in the Mackenzie giant radiating dyke swarm. *Nature* **356**, 511–513 (1992).
- Hargraves, R. B., Johnson, D. & Chen, C. Y. Distribution anisotropy: The cause of AMS in igneous rocks? *Geophys. Res. Lett.* **18**, 2193–2196 (1991).
- Richter, C., Kelso, P. R. & MacLeod, C. J. *Proc. ODP Sci. Res.* **147**, 393–403 (1996).
- Yaouancq, G. & MacLeod, C. J. Petrofabric investigation of gabbros from the Oman ophiolite: comparison between AMS and rock fabric. *Mar. Geophys. Res.* (in the press).
- Naslund, H. R. & McBirney, A. R. in *Layered Intrusions* (ed. Cawthorn, R. G.) 1–43 (Elsevier, Amsterdam, 1996).
- Benn, K. & Allard, B. Preferred mineral orientations related to magmatic flow in ophiolite layered gabbros. *J. Petrol.* **30**, 925–946 (1989).
- Staudigel, H., Gee, J., Tauxe, L. & Varga, R. J. Shallow intrusive directions of sheeted dikes in the Troodos ophiolite: anisotropy of magnetic susceptibility and structural data. *Geology* **20**, 841–844 (1992).
- Staudigel, H. *et al.* Geochemistry and intrusive directions in sheeted dikes in the Troodos ophiolite: implications for mid-ocean ridge spreading centers. *Geochem. Geophys. Geosys.* [online] **1**, 1999GC000001 (1999).
- Muller, M. R., Minshull, T. A. & White, R. S. Segmentation and melt supply at the South West Indian Ridge. *Geology* **27**, 867–870 (1999).
- Nicolas, A., Boudier, F. & Ildefonse, B. Variable crustal thickness in the Oman ophiolite: implication for oceanic crust. *J. Geophys. Res.* **101**, 17,941–17,950 (1996).
- Phipps-Morgan, J. & Parmentier, E. M. Crenulated seafloor: evidence for spreading-rate dependent structure of mantle upwelling and melting beneath a mid-oceanic spreading center. *Earth Planet. Sci. Lett.* **129**, 73–84 (1995).
- Mineral Resources Map of Cyprus* (Geological Survey Dept, Ministry of Agriculture and Natural Resources of Cyprus, Nicosia, 1982).
- Tauxe, L. *Paleomagnetic Principles and Practice* (Kluwer Academic, Dordrecht, 1998).
- Malpas, J. & Brace, T. *Geological Map of the Amiantos-Palekhori Area* (Geological Survey Dept, Ministry of Agriculture and Natural Resources of Cyprus, Nicosia, Cyprus, and Center of Earth Resources Research, Memorial University of Newfoundland, Canada, 1987).
- Jelinek, V. Characterization of the magnetic fabric of rocks. *Tectonophysics* **79**, T63–T67 (1981).

Acknowledgements

We thank A. Nicolas and J. Gee for comments and suggestions, and L. Tauxe for assistance with the statistical analysis. This research was supported by the US-Israel Binational Science Foundation.

Correspondence and requests for materials should be addressed to M.A. (e-mail: meira@mail.gsi.gov.il).

The Earth's 'missing' niobium may be in the core

J. Wade & B. J. Wood

Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK

As the Earth's metallic core segregated from the silicate mantle, some of the moderately siderophile ('iron-loving') elements such as vanadium and chromium^{1,2} are thought to have entered the metal phase, thus causing the observed depletions of these elements in the silicate part of the Earth. In contrast, refractory 'lithophile' elements such as calcium, scandium and the rare-earth elements are known to be present in the same proportions in the silicate portion of the Earth as in the chondritic meteorites—thought to represent primitive planetary material^{1,3}. Hence these lithophile elements apparently did not enter the core. Niobium has always been considered to be lithophile and refractory yet it has been observed to be depleted relative to other elements of the same type in the crust and upper mantle^{4,5}. This observation has been used to infer the existence of hidden niobium-rich reservoirs in the Earth's deep mantle⁵. Here we show, however, that niobium and vanadium partition in virtually identical fashion between liquid metal and liquid silicate at high pressure. Thus, if a significant fraction of the Earth's vanadium entered the core (as is thought), then so has a similar fraction of its niobium, and no hidden reservoir need be sought in the Earth's deep mantle.

Estimates of the bulk composition of the Earth^{1,2} are generally based on the compositions of the chondritic meteorites with which the Earth has strong chemical affinities. As the relative abundances of refractory lithophile elements are constant across the classes of these meteorites, their ratios (for example, Nb/Ta, Nb/La) in the bulk Earth can be estimated with confidence (Table 1). Although the relationship of these elements to refractory siderophile elements such as Fe, Ni and W is less certain, the mass of the Earth's core and its probable Fe/Ni ratio suggest that the bulk Earth has siderophile/lithophile ratios between those of CI and CM carbonaceous chondrites^{1,2} (Table 1). Core formation depleted the silicate mantle in the siderophile elements, but left non-siderophile elements in the same ratios one to another as in the bulk Earth. Later separation of the continental crust, which is highly enriched in many lithophile trace elements (the rare-earth elements, K, U, Th for example) left a large depleted reservoir that approximates the upper-mantle source of mid-ocean ridge basalts⁴. Table 1 shows some ratios of siderophile/lithophile and lithophile/lithophile elements in the bulk Earth, depleted upper mantle and the continental crust. The imprint of core formation can be seen as a marked reduction of siderophile/lithophile element ratios in upper mantle and crust relative to the bulk Earth. No detailed mass balance is required in order to demonstrate that the sum of continental crust

Table 1 Element ratios in terrestrial reservoirs

	Bulk Earth*	Silicate Earth	
		Upper mantle†	Continental crust‡
Fe/Al	20	2.7	0.6
Ni/Al	1.1	0.083	0.0006
Cr/Al	0.29	0.11	0.001
V/Yb	348	185	65
Nb/Ta	17.6	15.5	12.3
Nb/La	1.01	0.91	0.48
Ta/La	0.057	0.059	0.039

* Bulk Earth ratios are based on CI chondrites except for Fe/Al, Ni/Al, Cr/Al which appear to lie between CI and CM chondrites^{1,2}.

† Upper mantle values are based on MORB for the incompatible elements Nb, Ta, La and fertile peridotite¹ for the remainder.

‡ Refs 5, 18.

plus upper mantle cannot produce the Fe/Al and Ni/Al ratios of the bulk Earth. Most of the Earth's Fe and Ni have clearly entered the core. V and Cr are also somewhat depleted in the silicate part of the Earth, implying that the core contains significant proportions of the Earth's inventories of these two elements. The alternative hypothesis, that V and Cr are depleted in the mantle because they are relatively volatile and did not accrete to the Earth⁶, seems less likely. Available data on volatility during solar condensation⁷ indicate that V is less volatile than Fe and Ni, two elements that are undepleted in the Earth.

Turning to the refractory lithophile elements, Nb, Ta and La, it is clear that no mixture of crust and depleted upper mantle can generate chondritic ratios of Nb/Ta and Nb/La. There must be another reservoir in the Earth with superchondritic Nb/Ta and Nb/La (ref. 5). In contrast, Ta—the 'geochemical twin' of Nb—appears to have an essentially chondritic ratio to La in the silicate Earth, implying that no additional reservoir is required for this element. Rudnick *et al.*⁵ have argued that the Nb-rich reservoir is subducted oceanic crust transformed into refractory rutile-bearing eclogite. The main difficulties with this hypothesis are first, that large volumes (up to 6% of the mantle) of such eclogite are required for mass balance, and second, that its superchondritic Nb/Ta is not seen in the ocean-island basalts (HIMU basalts) which appear to contain subducted oceanic crust in their source regions^{8,9}. Here we explore an alternative possibility, that Nb is a weakly siderophile element slightly depleted in the silicate Earth due to partial dissolution in the core.

Free energies of reactions between metals and metal oxides¹⁰ suggest that Nb should be as siderophile as V and substantially more siderophile than Si, an element considered by some to constitute nearly 8% of the core². The implication is that Nb should have partitioned into the core in a manner similar to V. The data refer to pure metals and oxides at atmospheric pressure, however, rather than to very high pressure equilibration between silicate in the mantle and Fe alloy in the core^{11,12}. Observed concentrations of Ni and Co in the mantle^{1,2} are too high to be explained by low-pressure (0–2 GPa) equilibrium between metal and silicate containing about 6% Fe (ref. 11). But high-pressure experiments^{11–14} show that, with increasing pressure, these two elements become less siderophile with respect to Fe, and imply that single-stage core–mantle equilibrium would be possible at pressures of 25–30 GPa (refs 11, 13), plausibly at the base of a 700-km-deep magma ocean. Clearly, therefore, investigation of the effect of high pressure on the silicate–metal partitioning behaviour of weakly siderophile elements such as V, Cr and Nb is necessary in order to understand their depletions in the mantle.

We have performed liquid-metal/liquid-silicate partitioning experiments at low (2.5 GPa) and high (25 GPa) pressure for Nb and other lithophile and siderophile elements under conditions where the metal contains between 0.07% and 16.4% Si. Varying the Si content of the metal enabled us to study the relative siderophile behaviour of this element, as well as providing a means to vary the oxygen fugacity (f_{O_2}) of the experiment. Experimental starting materials were equal proportions of synthetic basalt, reduced at

Table 2 Experimental results

Conditions	2,300 °C, 25 GPa								1,750 °C, 2.5 GPa			
	No 6		No 4		No 5		No 3		NB2		NB3	
	Value	±	Value	±	Value	±	Value	±	Value	±	Value	±
Log f_{O_2} *	−1.95	0.05	−3.58	0.29	3.17	0.21	−1.13	0.17	−3.76	0.16	−4.99	0.18
Silicate												
Na	1.63	0.04	0.32	0.07	0.37	0.04	1.12	0.23	1.17	0.26	0.84	0.24
Mg	17.10	0.40	18.69	1.15	19.19	1.63	9.65	1.66	17.97	3.53	22.85	3.09
S	0.01	0.01	0.05	0.01	0.04	0.02	0.06	0.03	0.44	0.09	1.95	0.41
K	0.12	0.02	0.01	0.00	0.10	0.02	0.47	0.56	0.19	0.04	0.14	0.03
Ca	6.32	0.09	5.18	0.72	5.15	0.38	7.82	3.87	7.56	1.50	5.55	1.37
Ti	1.02	0.05	1.09	0.15	1.99	0.31	0.34	0.08	0.38	0.06	0.29	0.05
Cr	0.29	0.01	0.03	0.01	0.15	0.02	0.14	0.05	0.03	0.01	0.02	0.01
Mn	1.29	0.03	0.14	0.01	0.62	0.03	1.01	0.16	1.12	0.08	0.83	0.17
Fe	3.60	0.19	0.59	0.15	0.94	0.24	7.38	0.94	0.50	0.04	0.13	0.03
Al	5.18	0.13	5.48	0.33	3.82	0.33	8.35	2.61	6.27	1.14	4.73	1.06
Si	18.73	0.29	19.52	0.87	19.76	1.08	19.34	1.50	18.00	0.27	18.25	0.25
V	0.55	0.01	0.06	0.01	0.31	0.02	—	—	—	—	—	—
Nb	0.88	0.12	0.01	0.01	0.28	0.04	2.14	0.79	0.82	0.17	0.01	0.00
Mo	0.02	0.01	0.03	0.01	0.03	0.01	—	—	—	—	—	—
Ta	0.60	0.09	0.06	0.01	0.85	0.08	—	—	—	—	—	—
Ni	0.02	0.00	0.01	0.00	0.02	0.01	0.02	0.01	—	—	—	—
Ga	0.13	0.01	0.03	0.01	0.04	0.01	—	—	—	—	—	—
Hf	—	—	—	—	—	—	—	—	2.77	0.49	2.45	0.49
Zr	—	—	4.27	0.82	2.64	0.73	0.02	0.02	—	—	—	—
O	42.78	0.25	44.10	0.36	43.92	0.47	42.51	1.21	42.97	0.69	44.46	0.42
Total	100.27	0.26	99.67	0.41	100.21	0.56	100.36	0.92	100.21	0.74	102.49	0.34
Metal												
S	3.84	1.14	3.48	0.19	2.61	1.26	3.98	1.30	2.60	0.20	0.63	0.75
V	0.21	0.08	1.19	0.06	0.64	0.10	0.00	0.00	0.01	0.00	0.01	0.00
Cr	0.39	0.05	1.97	0.05	1.05	0.09	0.01	0.01	0.18	0.01	0.28	0.01
Mn	0.68	0.06	2.65	0.04	1.92	0.12	0.05	0.02	0.37	0.18	0.89	1.09
Fe	83.59	1.29	62.16	0.43	77.34	1.51	94.58	1.28	92.01	0.44	82.53	1.71
Co	2.02	0.05	1.61	0.02	1.82	0.05	0.49	0.02	0.45	0.01	0.53	0.02
Ni	2.19	0.06	1.76	0.02	1.88	0.04	0.38	0.02	0.35	0.01	0.44	0.02
W	1.64	0.12	1.28	0.06	1.23	0.09	—	—	—	—	—	—
Ga	1.16	0.13	1.31	0.02	1.35	0.10	—	—	—	—	—	—
Si	0.32	0.04	16.40	0.19	6.89	0.60	0.07	0.08	1.37	0.02	10.37	0.25
Nb	0.41	0.22	1.96	0.20	1.19	0.54	0.01	0.01	2.47	0.14	3.70	0.09
Ti	0.01	0.01	0.16	0.02	0.02	0.03	0.01	0.01	0.01	0.01	0.02	0.01
Mo	1.94	0.31	1.43	0.06	1.27	0.25	—	—	—	—	—	—
Ta	0.00	0.00	1.41	0.20	0.15	0.16	—	—	—	—	—	—
Os	1.51	0.26	1.25	0.05	1.19	0.13	—	—	—	—	—	—
Total	99.92	0.25	100.02	0.40	100.56	0.34	99.59	0.44	99.82	0.21	99.40	0.89

Errors expressed as 2 standard errors. Oxygen in silicate by stoichiometry.

* f_{O_2} expressed relative to IW buffer (see text: positive errors only shown).

1,000 °C in a CO/CO₂ atmosphere at 2 log f_{O_2} units below the fayalite–magnetite–quartz buffer, and a metallic component containing appropriate proportions of Fe, FeS, Fe₈₃Si₁₇ and FeSi. These were intimately ground under acetone with the trace-element component consisting of metals (Ni, Mo, Os, W) and oxides (of V, Nb, Ga, Ta, Cr, Ti, Mn, Co). At 2.5 GPa the samples were contained in MgO capsules and the experiments performed in a piston–cylinder apparatus at 1,750 °C for 30 min. The 25-GPa experiments were performed using the multi-anvil apparatus at the University of Bayreuth. Samples were contained in MgO capsules and held at 2,300 °C for approximately 90 s. Experiment duration was limited by the rapid reaction of the silicate with the capsule, generating inviscid ultrabasic melts (Table 2) that occasionally escaped from the capsule. Temperature control in all experiments was via a W/Re thermocouple, placed directly above the capsule; quenching was achieved by turning off the power to the furnace. Analysis was performed using the JEOL 8600 electron microprobe at the University of Bristol using a range of metal, oxide and silicate standards and conditions of 20 kV accelerating voltage and 15 nA beam current. Microprobe data were corrected using standard procedures. Results are given in Table 2 and shown in Figs 1–2.

Figure 1 shows the partition coefficients $D_{\text{metal/sil}} (= [\text{wt\% in metal}]/[\text{wt\% in silicate}])$ plotted as a function of oxygen fugacity at 2.5 GPa. For each experiment the latter was calculated relative to the iron–wüstite (IW) equilibrium, $\text{Fe} + 0.5\text{O}_2 = \text{FeO}$, by assuming that the activity of Fe (a_{Fe}) in the metal was equal to its mole fraction, and that the activity of FeO in the silicate was equal to $\text{Fe}/(\text{Fe}+\text{Mg})$. In our earlier work^{15,16}, a_{FeO} was obtained from the FeO mole fraction of (Mg,Fe)O solid solution coexisting with metal. We found that the latter correlates well with the $\text{Fe}/(\text{Fe}+\text{Mg})$ ratio of the melt. Therefore, in the present study we used $\text{Fe}/(\text{Fe}+\text{Mg})$ of melt to estimate a_{FeO} . Figure 1 shows that V, Cr and Nb have similar partitioning behaviour at 2.5 GPa, and that they partition into the metal much more strongly than does Si. At the oxygen fugacity appropriate for core–mantle equilibrium, partition coefficients of 0.01–0.1 would be expected for V and Cr and values close to 10^{-4} for Si. Ta (not shown) has similar partitioning behaviour to Si under these conditions¹⁵. The main implications of the figure are that if V and Cr are partially dissolved in the core, then some Nb should also

be present but little or no Ta or Si.

Results at 25 GPa (Fig. 2) are appropriate for investigating the hypothesis of core segregation at the bottom of a 700-km-deep magma ocean^{11–13}. Nb and V remain more siderophile than Si and, as at 2.5 GPa, have virtually identical metal–silicate partition coefficients to one another. In the oxygen fugacity range appropriate for core–mantle equilibrium, partition coefficients of 0.06–0.6 are found for these elements. Partition coefficients increase from 2.5 to 25 GPa and they probably also depend on temperature, so we cannot be sure that we have reproduced the partition coefficient which actually applied during core formation. However a D value of 0.6 would result in 23% of the element having entered the core,

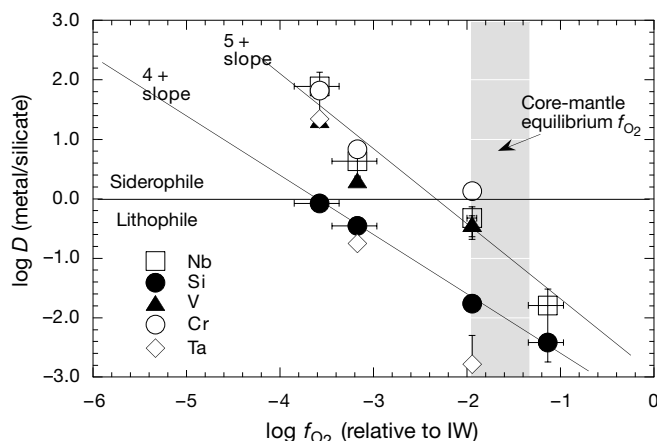


Figure 2 High-pressure data. The figure shows data from this study on partitioning between Fe-rich metallic liquid and silicate melt at 2,300 °C and 25 GPa. Partition coefficients and oxygen fugacities are calculated as in Fig. 1. Nb and V have almost identical tendencies to partition into the core, while Cr is slightly more siderophile and Ta much more lithophile. Nb data are consistent with 5+ oxidation state in the silicate. Apparent convergence of the behaviour of Ta, Nb, Cr and V at lowest f_{O_2} is probably an artefact due to changing Henry's law constants of the elements as the metal becomes very rich in Si (16.4%). Some f_{O_2} error bars have been omitted for clarity. At each f_{O_2} , error bars are same on all points.

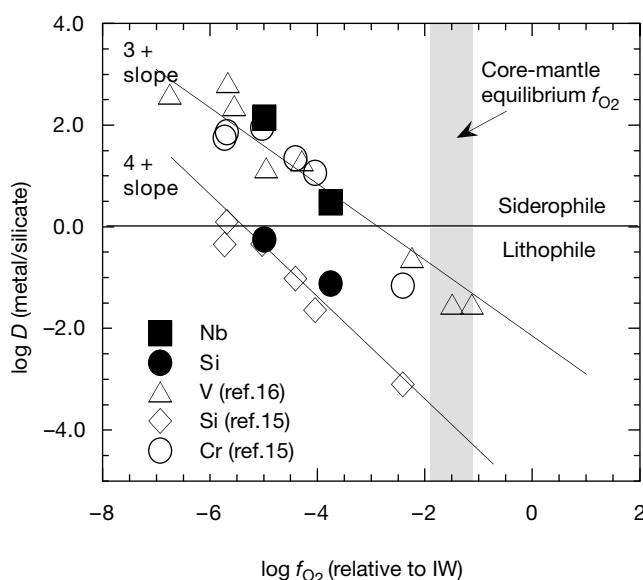


Figure 1 Low-pressure data. The figure shows partitioning between Fe-rich metallic liquid and silicate melt at 1,750 °C and 2.5 GPa. Solid symbols, data from this study; open symbols, data from refs 15 and 16. Partition coefficients D (metal/silicate) are computed on a weight basis, and oxygen fugacity (f_{O_2}) is expressed relative to the Fe–FeO buffer (IW; see text). Error bars are approximately the same size as the symbols. Lines through data

indicate that V and Cr are present in the 3+ oxidation state in the silicate, while Si is 4+. Shaded region, range of f_{O_2} appropriate for core–mantle equilibrium, calculated by assuming metal of 90% Fe coexisting with lower-mantle (Mg,Fe)O of composition $(\text{Fe}/\text{Fe}+\text{Mg}) = 0.1–0.2$ (ref. 19). Horizontal line at $D = 10^0$, conditions at which elements change from preferring silicate (lithophile) to metal (siderophile).

sufficient to produce observably subchondritic ratios of Nb/Ta and Nb/La in the silicate Earth. More robustly, the similar behaviour of V and Nb at both low and high pressure means that, if a significant proportion of the Earth's V is in the core, it must be accompanied by a similar fraction of its Nb. At the appropriate oxygen fugacity for single-stage core formation in the Earth, Cr is slightly more siderophile than Nb while Ta is much less siderophile than Nb (Fig. 2). These results are all consistent with significant dissolution of V, Cr and Nb in the core and the completely lithophile behaviour of Ta.

Partition coefficients for Si and Ga (Fig. 2, Table 2) suggest that Si is not a strong enough siderophile, even at 25 GPa, for the core to contain 8% Si and that Ga is a stronger siderophile than is required to explain its depletion in the silicate Earth¹. Temperature effects on $D_{\text{metal/sil}}$ can be large, however¹⁷, so the limited temperature range of our experiments preclude definitive conclusions in these cases.

The two competing hypotheses for Nb depletion in the silicate Earth depend only on the nature of the hidden reservoir. Either the core contains significant fractions of the Earth's V, Cr and Nb, or the depletions of V and Cr in the mantle are solely due to incomplete accretion to the Earth. In the latter case the Nb depletion must be due to a hidden silicate reservoir such as subducted refractory eclogite⁵. The second hypothesis would be supported if some HIMU basalts were found to have superchondritic Nb/Ta and Nb/La (ref. 5), or if V were found to be more volatile in the solar nebula than Fe. Current data suggest that neither of these conditions are met^{7,9}. □

Received 27 March; accepted 7 November 2000.

- McDonough, W. F. & Sun, S.-s. The composition of the Earth. *Chem. Geol.* **120**, 223–253 (1995).
- Allègre, C. J., Poirier, J.-P., Hummler, E. & Hofmann, A. W. The chemical composition of the Earth. *Earth Planet. Sci. Lett.* **134**, 515–526 (1995).
- Newsom, H. E. in *Global Earth Physics* (ed. Ahrens, T. J.) 159–189 (American Geophysical Union Reference Shelf 1, Washington DC, 1995).
- Hofmann, A. W. Chemical differentiation of the Earth: the relationship between mantle, continental crust and oceanic crust. *Earth Planet. Sci. Lett.* **90**, 297–314 (1988).
- Rudnick, R. L., Barth, M., Horn, I. & McDonough, W. F. Rutile-bearing refractory eclogites: missing link between continents and depleted mantle. *Science* **287**, 278–281 (2000).
- Drake, M. J., Newsom, H. E. & Capobianco, C. J. V, Cr and Mn in the Earth, Moon, EPB and SPB and the origin of the Moon: Experimental studies. *Geochim. Cosmochim. Acta* **53**, 2101–2111 (1989).
- Wasson, J. T. *Meteorites: Their Record of Early Solar-system History* (Freeman & Co., New York, 1995).
- Hofmann, A. W. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–228 (1997).
- Hofmann, A. W. & Jochum, K. P. Source characteristics derived from very incompatible trace elements in Mauna Loa and Mauna Kea basalts, Hawaii Scientific Drilling Project. *J. Geophys. Res.* **101**, 11831–11839 (1996).
- Robie, R. A., Hemingway, B. S. & Fisher, J. R. Thermodynamic properties of minerals and related substances at 298.15K and 1 bar (10^5 Pascals) pressure and at higher temperatures. *US Geol. Surv. Bull.* **1452** (1978).
- Li, J. & Agee, C. B. Geochemistry of mantle-core differentiation at high pressure. *Nature* **381**, 686–689 (1996).
- Righter, K., Drake, M. J. & Yaxley, G. Prediction of siderophile element metal-silicate partition coefficients to 20 GPa and 2800 degrees C: The effects of pressure, temperature, oxygen fugacity, and silicate and metallic melt compositions. *Phys. Earth Planet. Int.* **100**, 115–134 (1997).
- Righter, K. & Drake, M. J. Effect of water on metal-silicate partitioning of siderophile elements: a high pressure and temperature terrestrial magma ocean and core formation. *Earth Planet. Sci. Lett.* **171**, 383–399 (1999).
- Thibault, Y. & Walter, M. J. The influence of pressure and temperature on the metal-silicate partition-coefficients of nickel and cobalt in a model-c1 chondrite and implications for metal segregation in a deep magma ocean. *Geochim. Cosmochim. Acta* **59**, 991–1002 (1995).
- Kilburn, M. R. & Wood, B. J. Metal-silicate partitioning and the incompatibility of S and Si during core formation. *Earth Planet. Sci. Lett.* **152**, 139–148 (1997).
- Kilburn, M. R. *Geochemical Constraints on the Formation of the Earth's Core*. Thesis, Univ. Bristol (1999).
- Gessmann, C. K., Wood, B. J., Rubie, D. C. & Kilburn, M. R. Solubility of silicon in liquid metal at high pressure: implications for the composition of the Earth's core. *Earth Planet. Sci. Lett.* (in the press).
- Rudnick, R. L. Making continental crust. *Nature* **378**, 571–578 (1995).
- Wood, B. J. Phase transformations and partitioning relations in peridotite under lower mantle conditions. *Earth Planet. Sci. Lett.* **174**, 341–354 (2000).

Acknowledgements

This work was supported by the NERC. Experiments at Bayreuth were performed with assistance from the EU Large Scale Facility programme. B.J.W. acknowledges a Max Planck research award.

Correspondence and requests for materials should be addressed to B.J.W. (e-mail: b.j.wood@bris.ac.uk).

Symbiotic fungal endophytes control insect host–parasite interaction webs

Marina Omacini*, Enrique J. Chaneton*, Claudio M. Ghera* & Christine B. Müller†‡

* IFEVA-Departamento de Recursos Naturales y Ambiente, Facultad de Agronomía, Universidad de Buenos Aires, Av. San Martín 4453, 1417 Buenos Aires, Argentina

† Department of Biology and NERC Centre for Population Biology, Imperial College at Silwood Park, Ascot, Berkshire SL5 7PY, UK

‡ The Zoological Society of London, Institute of Zoology, Regent's Park, London NW1 4RY, UK

Symbiotic microorganisms that live intimately associated with terrestrial plants affect both the quantity and quality of resources^{1,2}, and thus the energy supply to consumer populations at higher levels in the food chain. Empirical evidence on resource limitation of food webs points to primary productivity as a major determinant of consumer abundance and trophic structure^{3–6}. Prey quality plays a critical role in community regulation^{7,8}. Plants infected by endophytic fungi are known to be chemically protected against herbivore consumption^{9–11}. However, the influence of this microbe–plant association on multi-trophic interactions remains largely unexplored. Here we present the effects of fungal endophytes on insect food webs that reflect limited energy transfer to consumers as a result of low plant quality, rather than low productivity. Herbivore–parasite webs on endophyte-free grasses show enhanced insect abundance at alternate trophic levels, higher rates of parasitism, and increased dominance by a few trophic links. These results mirror predicted effects of increased productivity on food-web dynamics¹². Thus 'hidden' microbial symbionts can have community-wide impacts on the pattern and strength of resource–consumer interactions.

A central issue in ecology is to understand to what extent populations are limited by resources or natural enemies^{13–16}. There is emerging consensus that the interplay between these forces control food-web structure in response to resource enrichment^{6,17} as well as across natural gradients of productivity^{7,12}. The potential role of symbiotic microorganisms, such as mycorrhizae and endophytes, in the regulation of terrestrial food webs has only recently been addressed¹⁵. Fungal endophytes in the genus *Neotyphodium* (Ascomycetes: Clavicipitaceae) form mutualistic associations with a variety of grasses^{9,11}. The fungal hyphae grow intercellularly in leaf and stem tissue, causing asymptomatic infections that are transmitted exclusively through the seeds of the host plant. Endophytic fungi obtain nutrients from their hosts, whereas infected plants may gain protection from insect herbivores or vertebrate grazers via the toxic or deterrent effects of alkaloids synthesized by the fungus^{9,10}. No study to date has examined the impact of these fungal symbionts on multi-trophic insect assemblages. We predicted that plant infection with endophytes would alter herbivore abundance and strength of interactions at higher trophic levels.

We studied the structure of an aphid–parasite food web, naturally assembled on monocultures of *Lolium multiflorum* (Italian ryegrass) grown from *Neotyphodium*-infected or endophyte-free seeds. Aphids (Homoptera: Aphididae) are external plant feeders that form species-rich communities with their hymenopteran parasitoids. Aphid parasitoids encompass both specialist and polyphagous parasitic insects, whose larvae develop singly within an aphid host, eventually killing it. Dead aphids turn into a 'mummy' in which parasitoid pupation takes place. Parasitoids can be sorted conveniently into trophic levels. 'Primary' parasitoids attack aphid

nymphal instars. 'Secondary' parasitoids attack and consume primary parasitoids, either within a living parasitized aphid (hyperparasitoids) or a mummified aphid (mummy parasitoids). Mummy parasitoids may also kill hyperparasitoids when they co-occur; they are therefore the top consumers in the food chain. Interactions between aphids and parasitoids are easy to quantify¹⁸ as only one adult primary or secondary parasitoid emerges from each mummy. When secondary parasitoids are present, they consume the primary parasitoid before emergence. In this case, secondary parasitoids in the interaction web are linked to the source aphid¹⁸. We used this system to examine the impact of fungal endophytes on a food chain of four trophic levels.

The density of aphid herbivores was three times higher on endophyte-free grass monocultures (−E) than on endophyte-infected plots (+E) (Fig. 1a). The two aphid species colonizing the experiment, *Rhopalosiphum padi* and *Metopolophium festucae*, were differently affected by the endophyte treatment (Fig. 1a). Removing the endophyte also enhanced parasitoid activity, resulting in an 8-fold increase in total density of parasitized aphids ($t_{18} = 4.27$, $P < 0.001$). This effect was driven by the higher number of *R. padi* mummies on −E plots ($t_{18} = 2.91$, $P < 0.01$; for *M. festucae*, $P = 0.35$). More importantly, aphids suffered a proportionally higher rate of parasitism on −E plots than on +E ones (Fig. 1b). Despite this overall increase in parasitoid pressure, primary parasitoids were relatively less successful on endophyte-free plots, as a result of a disproportionate increase in secondary parasitism (secondary/−

primary parasitoid ratio: +E = 2.1 versus −E = 6.5, $\chi^2 = 4.19$, $P = 0.041$; Fig. 1b). In addition, fungal endophytes influenced individual parasitoid traits, without affecting aphid mummy size ($F_{1,87} = 0.49$). The body size of secondary parasitoids emerging from *R. padi* mummies on −E plots was larger than of those from mummies on +E plots (two-way analysis of covariance, ANCOVA, endophyte effect: $F_{1,41} = 5.56$, $P = 0.023$, endophyte $\times$ sex: $P = 0.45$, host size covariate: $F_{1,41} = 3.28$, $P = 0.08$). No effect of endophyte infection on body size was detected for secondary parasitoids emerging from *M. festucae* ($P = 0.82$). Thus, removing the limitation imposed by endophytes on herbivore density generated a cascade of effects, with unequal consequences for consumers at upper trophic levels. Whereas herbivores and top parasites

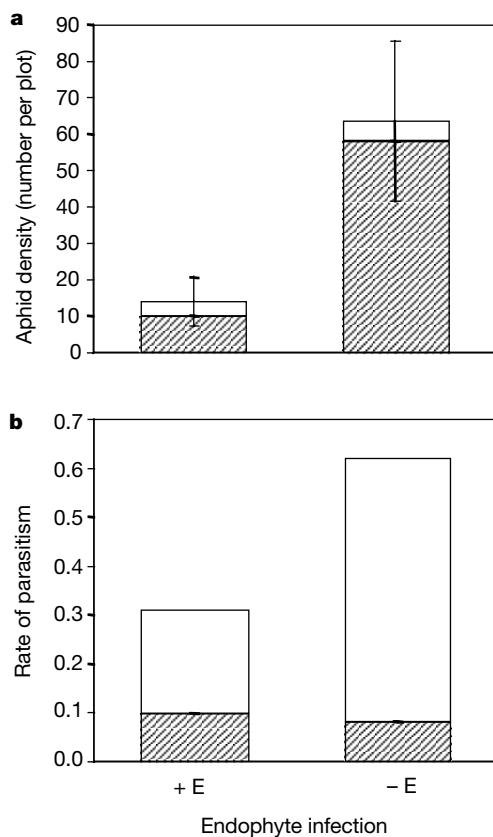


Figure 1 Response of insects to the grass-endophyte association. **a**, Density response of the aphids *Rhopalosiphum padi* (shaded bars) and *Metopolophium festucae* (empty bars) to the removal of fungal endophytes. Bars show that the mean ($\pm$ s.e.m.) total aphid density differed ($t_{18} = 2.42$, $P = 0.026$) between plots with (+E) and without (−E) endophytes. This resulted from a significant difference for *R. padi* ($t_{18} = 3.00$, $P = 0.008$) but not for *M. festucae* ($P = 0.63$). **b**, Total rate of aphid parasitism in the two endophyte treatments, including the proportion of emerged primary (shaded bars) and secondary (open bars) parasitoids. The rate of parasitism reflects the proportion of all aphids that were mummified; a randomization test showed that this rate differed between treatments ($P < 0.025$).

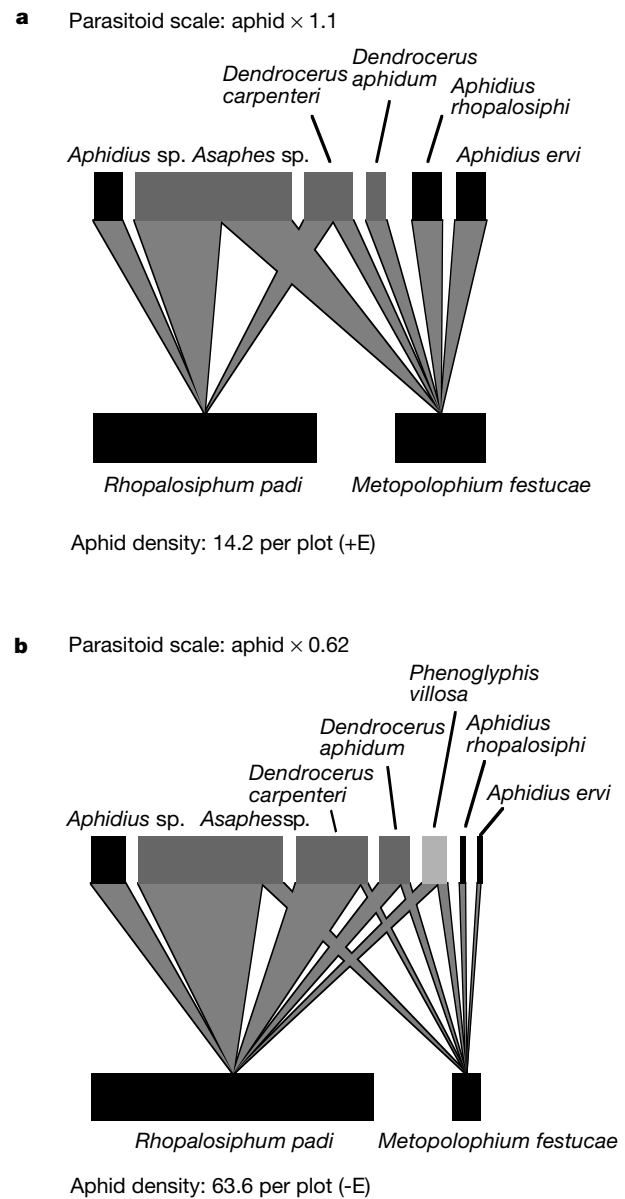


Figure 2 Aphid–parasite interaction webs established on *Lolium multiflorum* monocultures. **a**, With (+E) and **b**, without (−E) infection by fungal endophytes. The length of the horizontal bars represents the population density of two aphid species (bottom level) and their hymenopteran parasitoid consumers (top level), including primary parasitoids (black bars), secondary mummy parasitoids (dark grey bars) and a hyperparasitoid (light grey bar). The basal width of the connections from parasitoids to aphids reflects the proportion of different parasitoid species that emerged from each host species per treatment. Web diagrams are scaled according to the respective mean aphid densities to aid visual comparison of treatments.

reacted positively to endophyte removal, intermediate parasites showed no substantial change in reproductive success, despite their increased attack rate. Such 'bottom-up cascades' are predicted by food-chain models that stress the interaction between basal resources and dynamical consumer feedbacks in the control of community structure^{3,12,15}.

Our food web was structured upon two aphid-based trophic chains that were connected at the top level by polyphagous secondary parasitoids (Fig. 2). The increased dominance of *R. padi* on -E plots produced an increase in the median strength of parasitic interactions (Table 1). Conversely, trophic links between parasitoids and *M. festucae* became weaker compared to those based on *R. padi* (Fig. 2b), which reduced the evenness of pairwise interaction strengths across the web (Table 1). Endophyte removal enhanced the complexity of the host-parasite web by increasing the number of trophic links per species, connectance, and the number of indirect links through shared parasitoids (Table 1; Fig. 2). This was partly explained by the appearance of a generalist hyperparasitoid on -E plots (Fig. 2b). Moreover, the addition of a hyperparasitoid species effectively generated a longer food chain on endophyte-free grass monocultures. The observed changes in food-web structure were driven mainly by the different response of the two aphid species to the endophyte treatment. Several mechanisms may account for this heterogeneous reaction within the herbivore trophic level. First, the two aphid species may have different sensitivity to endophyte infection¹⁰. Second, *R. padi* may have a competitive advantage over *M. festucae* that is amplified in the absence of endophytes. Third, the larger number of parasitoids supported by *R. padi* on -E plots could increase the consumer pressure on *M. festucae* through parasite-mediated apparent competition¹⁹⁻²¹. Theoretical models suggest that the existence of weak trophic links maintained by generalist consumers is important in promoting community persistence and stability²².

We initially assumed that endophyte effects on insect consumers would be mediated by measurable changes in productivity and/or quality of the plant resource. Here, changes in insect performance occurred without significant differences in total above-ground plant biomass between +E (210.7 g) and -E (191.6 g per 0.25 m²) monocultures ($t_8 = 0.04$, $P = 0.97$). In addition, we measured leaf nitrogen (N) concentration as an estimate of endophyte effects on plant quality. Against expectation, endophyte removal decreased leaf N concentration in *L. multiflorum* (+E = 1.02 ± 0.04 versus -E = 0.89 ± 0.01 , $t_8 = 3.47$, $P < 0.01$). This result might indicate that part of the extra N-infected plants is diverted into alkaloid synthesis by the fungus, making it inaccessible to insect herbivores¹⁰. Together with the well-documented accumulation of chemical defences in endophyte-infected grasses⁹⁻¹¹, our results suggest that consumers were limited by unidentified components of resource quality, but not by primary productivity.

It is conceivable that the effect of plant endosymbionts on food webs will cascade up through various trophic pathways. We found

that the frequency of grass stems attacked by leaf mining insects was consistently higher on -E (33.3%) than on +E plots (19.3%) throughout the experiment (repeated-measures analysis of variance ANOVA, endophyte effect: $F_{1,18} = 7.83$, $P = 0.012$; endophyte $\times$ date: $F_{2,36} = 0.16$, $P = 0.86$). Negative impacts of fungal endophytes on different phytophagous insects were reported by many other studies⁹⁻¹¹. Moreover, a few laboratory experiments showed that grass endophytes can affect parasitoid performance^{23,24}. Our experiment demonstrates for the first time that endosymbionts of plants are able to alter species abundance and consumer-resource interactions across multiple trophic levels in a field setting.

Microorganisms can greatly affect many terrestrial communities²⁵. Research on the ecological role of microbial symbionts has focused on their impact in plant communities^{2,26}. In particular, endophytes can mediate competitive interactions between plant species affecting vegetation diversity and succession²⁷. We have shown that fungal endophytes control food-web structure by disrupting the transfer of energy from plants to upper trophic levels. As endophytes live concealed within the host plant tissue, their impact on natural communities and biodiversity may easily be overlooked. These inconspicuous mutualistic associations can, however, exert a regulatory force on food-web dynamics that is qualitatively similar to that of primary productivity or nutrient supply^{5-7,17} in many ecosystems. □

Methods

Experimental design

Forty plots of Italian ryegrass were established on 2 July 1999, using 50 $\times$ 50 $\times$ 15 cm wooden containers laid out in an experimental garden at the College of Agronomy, University of Buenos Aires. Plots were arranged in a 5 $\times$ 8 grid, with 30-cm-wide corridors kept short by mowing. The containers were filled with soil extracted from a grassland site in the Inland Pampa, eastern Argentina. The seeds for the experiment were collected in the same area from old-field communities dominated by *L. multiflorum* populations with high levels (85–95%) of endophyte infection. Before the experiment, a batch of seeds was treated with a systemic fungicide (triadimenol: 5 mg per g seed) to obtain endophyte-free plants. Plots were randomly assigned to one of two treatments, with (+E) or without (-E) endophyte infection, and were sown with 2 g of *L. multiflorum* seed (~700 seeds per plot). These grass monocultures were maintained by frequent weeding. Levels of endophyte infection were confirmed at the end of the experiment by microscopic examination (aniline blue-lactic acid stain) of 30 seeds taken from ten plants per plot. The fungal mycelium was found to be associated to the aleurone layer of the seed²⁸. 95% of seeds collected from +E plots were infected, whereas seeds collected from -E plots contained no fungal endophytes.

Insect and plant sampling

Six months after sowing, at peak above-ground grass biomass, the natural occurrence of herbivorous insects was recorded in 10 plots of each treatment. Three times between 4 and 30 November 1999, the density of aphids and their hymenopteran parasitoids was estimated by counting the number of living aphids and mummies on 40 grass stems selected at random within each plot. The assessment of the aphid-parasite interaction web was based upon counts made on 4 November when aphid and parasitoid abundances were at their observed maximum. Thus, we report effects on insect densities that most probably integrated population responses across several aphid generations, followed by the aggregative response of parasitoids to differences in prey availability. Insect densities were compared between +E and -E grass plots using *t*-tests on log-transformed data. Differences in the total rate of parasitism were evaluated through a randomization test, using Resampling Stats version 4.2 (Julian Simon, Resampling Stats, Inc.). The contribution of primary and secondary parasitoids to overall parasitoid emergence from mummified aphids was examined with the χ^2 statistic.

To produce an enhanced description of the parasitoid community, the sample of mummified aphids obtained during the censuses was supplemented with an extensive collection of mummies from across all the plots. Parasitoids were reared individually in gelatine capsules under ambient temperature and identified and measured (body size: length in mm) upon emergence. Endophyte-driven effects on parasitoid body size were tested using factorial ANCOVA, with endophyte treatment and parasitoid sex as the main effects, and mummy size as a covariate.

Plant density was estimated in late November by counting the number of grass stems within a 6 $\times$ 30 cm strip-quadrat in each plot. This measure was used to adjust insect densities to a common, plot-area basis¹⁸. To determine above-ground plant biomass (g dry weight, after 48 h at 72 °C), a grass sample cut to soil level within a 10-cm-diameter cylinder was extracted from five replicate plots per treatment at the end of the experiment (December). Plant quality was assessed by measuring total leaf nitrogen concentration in a composite sample of five plants per plot, using a semi-micro Kjeldahl acid digestion.

Interaction webs

The webs were constructed using mean densities of living and mummified aphids to

Table 1 Effect of fungal endophyte infection on insect host-parasitoid interaction webs naturally established on *Lolium multiflorum* grass monocultures

Food-web attributes	Endophyte	
	Present	Absent
Total species richness	8	9
Linkage density	1.0	1.2
Connectance	0.67 (8/12)	0.79 (11/14)
Shared parasitism	0.33 (2/6)	0.57 (4/7)
Interaction strength		
Median	1.18	2.83
Evenness	0.84	0.62

Linkage density reflects the number of observed aphid-parasitoid links divided by the total number of species in the web. Connectance refers to the ratio between observed and maximum possible number of trophic links in a web. Shared parasitism is the proportion of parasitoid species attacking both aphid hosts. For descriptors of interaction strength, *P* values were derived from rank tests.

obtain a quantitative description of herbivore and parasitoid trophic levels at the scale of plots of 0.25 m². The strength of each pairwise parasitic interaction within a treatment was estimated by determining the proportional contribution of each parasitoid species to the total number of parasites emerged from a given aphid species in laboratory rearings¹⁸. These proportional estimates of parasitoid abundance per host species and treatment were translated into absolute measures of interaction strength after multiplying by the mean density of parasitized aphids (that is, mummy density per host species) recorded in the field censuses. We determined levels of linkage density, connectance, and shared parasitism for the food web representing each separate treatment. The median strength of all pairwise aphid–parasite interactions was computed for each food web and compared by the Mann–Whitney two-sample test. The distribution of parasitic interaction strengths within each web was summarized using a standard evenness index²⁹ and was evaluated statistically after jackknife resampling of the original data³⁰.

Received 21 September; accepted 31 October 2000.

- Gehring, C. A. & Whitham, T. G. Interactions between aboveground herbivores and mycorrhizal mutualists of plants. *Trends Ecol. Evol.* **9**, 251–255 (1994).
- van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69–72 (1998).
- Wootton, J. T. & Power, M. E. Productivity, consumers, and the structure of a river food chain. *Proc. Natl Acad. Sci. USA* **90**, 1384–1387 (1993).
- Kaunzinger, C. M. K. & Morin, P. J. Productivity controls food-chain properties in microbial communities. *Nature* **395**, 495–497 (1998).
- Mikola, J. & Setälä, H. Productivity and trophic-level biomasses in a microbial-based solid food web. *Oikos* **82**, 158–168 (1998).
- Hulot, F. D., Lacroix, G., Lescher-Moutoué, F. & Loreau, M. Functional diversity governs ecosystem response to nutrient enrichment. *Nature* **405**, 340–344 (2000).
- Leibold, M. A., Chase, J. M., Schurin, J. B. & Downing, A. L. Species turnover and the regulation of trophic structure. *Annu. Rev. Ecol. Syst.* **28**, 467–494 (1997).
- Abrams, P. A. Effects of increased productivity on the abundance of trophic levels. *Am. Nat.* **141**, 351–371 (1993).
- Clay, K. Fungal endophytes of grasses. *Annu. Rev. Ecol. Syst.* **21**, 275–297 (1990).
- Dahlman, D. L., Eichensoer, H. & Siegel, M. R. in *Microbial Mediation of Plant–Herbivore Interactions* (eds Barbosa, P., Kirschik, V. A. & Jones, C. G.) 227–252 (John Wiley, New York, 1991).
- Breen, J. P. *Acremonium* endophyte interactions with enhanced plant resistance to insects. *Annu. Rev. Entomol.* **39**, 401–423 (1994).
- Oksanen, L., Fretwell, S. D., Arruda, J. & Niemelä, P. Exploitation ecosystems in gradients of primary productivity. *Am. Nat.* **118**, 240–261 (1981).
- Hairton, N. G., Smith, F. E. & Slobotkin, L. B. Community structure, population control, and competition. *Am. Nat.* **94**, 421–425 (1960).
- Price, P. W. *et al.* Interaction among three trophic levels: influence of plants on interactions between insect herbivores and natural enemies. *Annu. Rev. Ecol. Syst.* **11**, 41–65 (1980).
- Hunter, M. D. & Price, P. W. Playing chutes and ladders: heterogeneity and the relative roles of bottom-up and top-down forces in natural communities. *Ecology* **73**, 724–732 (1992).
- Osenberg, C. W. & Mittelbach, G. G. in *Food Webs: Integration of Patterns and Dynamics* (eds Polis, G. & Winemiller, K. O.) 134–148 (Chapman & Hall, New York, 1996).
- DeAngelis, D. L. *Dynamics of Nutrient Cycling and Food Webs* (Chapman & Hall, London, 1992).
- Müller, C. B., Adriaanse, I. C. T., Belshaw, R. & Godfray, H. C. J. The structure of an aphid–parasitoid community. *J. Anim. Ecol.* **68**, 346–370 (1999).
- Holt, R. D. Predation, apparent competition and the structure of prey communities. *Theor. Pop. Biol.* **12**, 197–229 (1977).
- Holt, R. D. & Lawton, J. H. Apparent competition and enemy-free space in insect host–parasitoid communities. *Am. Nat.* **142**, 623–645 (1993).
- Chanton, E. J. & Bonsall, M. B. Enemy-mediated apparent competition: empirical patterns and the evidence. *Oikos* **88**, 380–394 (2000).
- McCann, K., Hastings, A. & Huxel, G. Weak trophic interactions and the balance of nature. *Nature* **395**, 794–798 (1998).
- Barker, G. M. & Addison, P. J. Influence of Clavicipitaceous endophyte infection in ryegrass on development of the parasitoid *Microctonus hyperodae* Loeb (Hymenoptera: Braconidae) in *Listronotus bonariensis* (Kuschel) (Coleoptera: Curculionidae). *Biol. Contr.* **7**, 281–287 (1996).
- Bultman, T. L., Borowicz, K. L., Schneble, R. M., Coudron, T. A. & Bush, L. P. Effect of a fungal endophyte on the growth and survival of two *Euplectrus* parasitoids. *Oikos* **78**, 170–176 (1997).
- Dobson, A. P. & Crawley, M. J. Pathogens and the structure of plant communities. *Trends Ecol. Evol.* **9**, 393–398 (1994).
- Grime, J. P., MacKey, J. M. L., Hillier, S. H. & Read, D. J. Floristic diversity in a model system using experimental microcosms. *Nature* **328**, 420–422 (1987).
- Clay, K. & Holah, J. Fungal endophyte symbiosis and plant diversity in successional fields. *Science* **285**, 1742–1744 (1999).
- Latch, G. C. M., Christensen, M. J. & Hickson, R. E. Endophytes of annual and hybrid ryegrasses. *New Zeal. J. Agr. Res.* **31**, 57–63 (1988).
- Alatalo, R. V. Problems in the measurement of evenness in ecology. *Oikos* **37**, 199–204 (1981).
- Magurran, A. E. *Ecological Diversity and its Measurement* (Princeton Univ. Press, New Jersey, 1988).

Acknowledgements

We thank A. Austin, M. Bonsall, A. Bourke, C. Godfray, D. Golombek, T. H. Jones, N. Mazia, R. Pettifor, S. Power, P. Roset, S. Semple and M. Vila-Aiub for comments on the manuscript; E. Demartin, P. Gundel and M. Rabadan for field assistance; R. Belshaw and F. van Veen for helping with parasitoid identification; and C. Godfray for constructing the web diagrams. This study was funded by grants from the Agencia Nacional de Promoción Científica y Tecnológica de Argentina and Fundación Antorchas.

Correspondence and requests for materials should be addressed to M.O. (e-mail: omacini@ifeva.edu.ar) or C.B.M. (e-mail: christine.mueller@ioz.ac.uk).

Evolution of the bilaterian larval foregut

Detlev Arendt*, Ulrich Technau† & Joachim Wittbrodt*

* European Molecular Biology Laboratory, Developmental Biology Programme, Meyerhofstrasse 1, 69012 Heidelberg, Germany

† Molecular cell biology, Zoological Institute, Darmstadt University of Technology, Schnittspahnstrasse 10, 64287 Darmstadt, Germany

Bilateria are subdivided into Protostomia and Deuterostomia^{1,2}. Indirect development through primary, ciliary larvae occurs in both of these branches; however, the closing blastopore develops into mouth and anus in Protostomia and into anus only in Deuterostomia. Because of this important difference in larval gut ontogeny, the tube-shaped guts in protostome and deuterostome primary larvae are thought to have evolved independently^{2,3}. To test this hypothesis, we have analysed the expression of *brachyury*, *otx* and *gooseoid* homologues in the polychaete

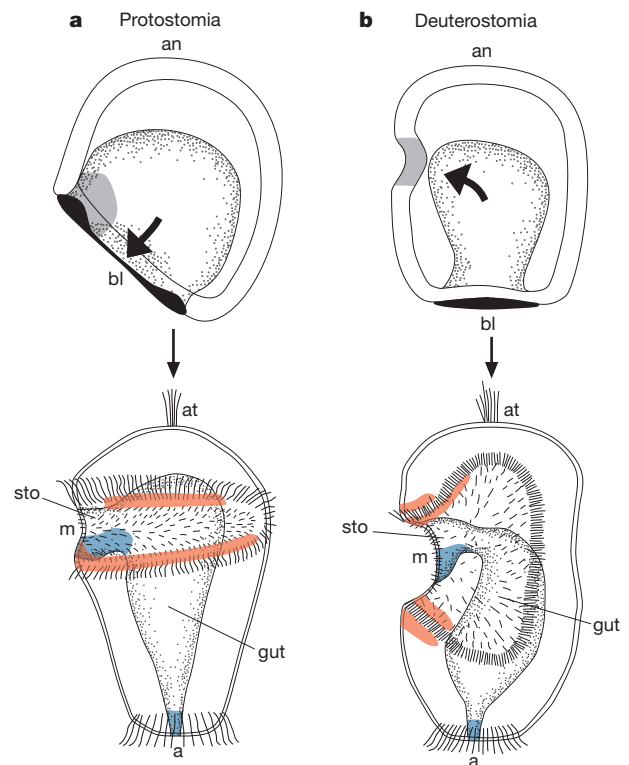


Figure 1 Different ontogeny but similar body plans of Protostomia and Deuterostomia primary larvae as shown by similar expression of *brachyury* in the ventral developing foregut and *otx* in ciliary bands bordering the mouth region. Late gastrula embryos (top) develop into pelagic, ciliary primary larvae (bottom). **a**, Polychaeta (Protostomia). The lateral blastopore lips fuse along the later ventral midline. The blastopore gives rise to mouth and anus at opposite ends. In the trochophore larva, *brachyury* (blue) is expressed in the ventral portion of the stomodaeum and in the proctodaeum, and *otx* (red) is expressed in two bands of cells along the preoral prototroch and the postoral metatroch. **b**, Enteropneusta (Deuterostomia). The tip of the gastrulation cavity touches the lateral body wall on the future ventral side, where the mouth later breaks through. The blastopore gives rise to the anus only. In the early tornaria larva, *brachyury* (blue) is expressed in the ventral portion of the stomodaeum and in the proctodaeum⁷, and *otx* (red) is expressed in two upper bands parallel to the preoral ciliated band and in two lower bands parallel to the postoral ciliated band⁹. A similar *otx* pattern is observed in the 30-h auricularia of the sea cucumber⁸. a, anus; an, animal pole; at, apical tuft; bl, blastopore; m, mouth; sto, stomodaeum. (Ciliary larvae adapted from ref. 2).

*Platynereis dumerilii*⁴, which develops by means of a trochophora larva—the primary, ciliary larva prototypic for Protostomia². Here we show that *brachyury* expression in the ventral portion of the developing foregut in *Platynereis* and also *otx* expression along ciliated bands in the mouth region of the trochophora larva parallels expression in primary larvae in Deuterostomia^{5–9}. In addition, *gooseoid* expression in the foregut of *Platynereis* mirrors the function in higher Deuterostomia¹⁰. We present molecular evidence for the evolutionary conservation of larval foreguts and mouth regions of Protostomia and Deuterostomia. Our data indicate that Urbilateria, the common bilaterian ancestors, developed through a primary, ciliary larva that already possessed a tripartite tube-shaped gut.

In the ciliary larvae of basal Deuterostomia (Echinodermata, Enteropneusta), the T-box gene *brachyury* has very restricted expression comprising the ventral portion of the developing foregut and the developing hindgut^{5–7} (Fig. 1). In contrast, in higher Protostomia, such as insects¹¹ and nematodes¹², *brachyury* is expressed in the hindgut but not in the foregut. Accordingly, if protostome and deuterostome larval foreguts evolved independently, *brachyury* expression should be restricted to the hindgut of the trochophora larva. We cloned the *Platynereis brachyury* homologue (*Pd-bra*) containing an amino-terminal T-box with diagnostic amino acids

typical for Brachyury¹³, which in phylogenetic analysis clearly clusters with members of the T/Brachyury subfamily of T-box proteins (Fig. 2a; and data not shown). During *Platynereis* embryogenesis, *Pd-bra* expression is first detected at 8 h of development in vegetal cells around the closing blastopore (data not shown). In the post-gastrula larva, *Pd-bra* demarcates the outline of the closed, slit-like blastopore (22 h, Fig. 2b, c). From anterior to posterior, the stained cells represent developing ventral foregut (stomodaeum), ventral midline cells and developing hindgut (proctodaeum) surrounded by a few posterior mesodermal cells. In the late trochophora larva, *Pd-bra* expression is restricted to the ventral portion of the developing foregut and to the developing hindgut (36 h, Fig. 2d). This closely resembles *brachyury* expression in the ciliary larvae of basal deuterostomes^{5–7,14} (Fig. 1, blue).

There is co-specification of hindgut and posterior mesoderm in *Drosophila*¹⁵ enteropneusts^{6,7} and vertebrates¹⁶. *Pd-bra* is also expressed in a small, distinct population of cells that ultimately gives rise to both hindgut and visceral mesoderm in nereid polychaetes¹⁷ from the earliest stages onwards (Fig 3b and d; and data not shown). This corroborates homology of hindguts throughout Bilateria, as proposed on the basis of shared expression of *caudal* homologues¹⁸. *Pd-bra* expression is also seen in an early, distinct population of ventral midline cells present in the polychaete

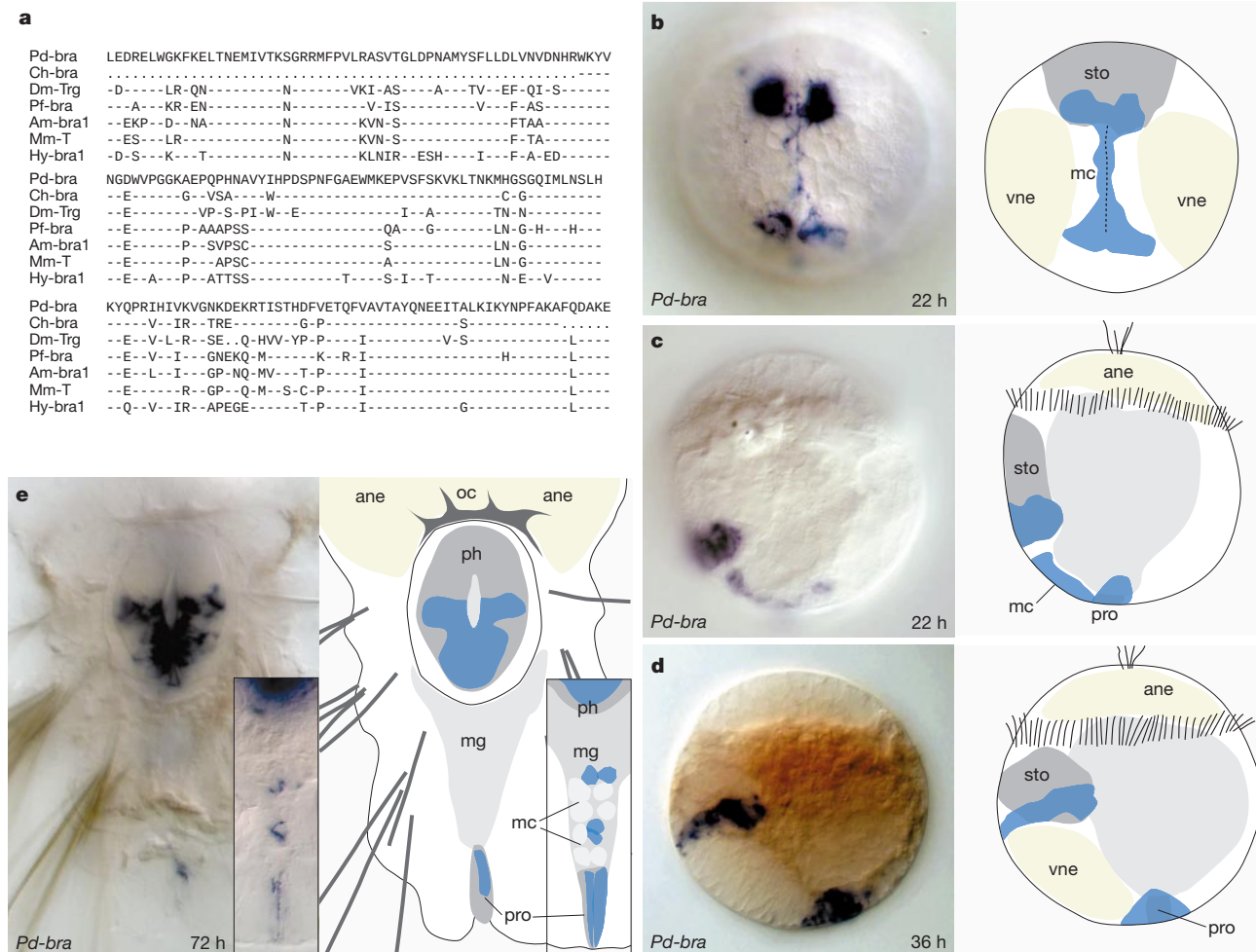


Figure 2 Expression of *Pd-bra*. Stomodaeum or proctodaeum anlage (dark-grey), yolk midgut anlage (light grey) and neuroectoderm (yellow) are shown; stippled line represents the ventral midline. **a**, Amino-acid sequence alignment of Brachyury T-box domains. **b**, Ventral–posterior view of 22-h trochophora where *Pd-bra* expression (blue) outlines the slit-like blastopore. **c**, Lateral view of 22-h trochophora. **d**, Lateral view of 36-h late trochophora with prominent expression in ventral stomodaeum, proctodaeum and

posterior mesoderm. **e**, Ventral view of 72-h young worm with strong expression in ventral stomodaeum. Inset shows a more ventral focal plane. Note the expression in midline cells. ane, apical neuroectoderm; mc, midline cells; mg, midgut; ph, pharynx; pm, posterior mesoderm; pro, proctodaeum; sto, stomodaeum; vne, ventral neuroectoderm; Pd, *Platynereis dumerilii*; Ch, *Chaetopterus spec.*; Dm, *Drosophila melanogaster*; Pf, *Ptychodera flava*; Am, *Branchiostoma floridae*; Mm, *Mus musculus*; Hy, *Hydra vulgaris*.

trochophora (Fig. 3b). Although *Pd-bra* is not detected in midline cells of 36 h or 48 h larvae, expression reappears along the midline of the developing young worm (72 h, Fig. 3e, inset). In cross-sections, *Pd-bra*-positive cells can be identified as a ventral-medial part of the cellularizing midgut, facing the developing nerve cord (data not shown). Notably, early contiguous *brachyury* expression between

later stomodaeum and proctodaeum also covers the ventral midline in starfish larva⁵. Furthermore, in chordates *brachyury* is expressed in dorsal midline cells that later give rise to notochord and that, by position, correspond to ventral midline cells in non-chordate groups^{19,20}. This makes *brachyury*-expressing midline cells another potentially conserved feature of the Bilateria.

To address homology of the larval mouth regions, we analysed the homeobox gene *otx/orthodenticle*. In the ciliary larvae of basal Deuterostomia, *otx* homologues are expressed specifically along ciliary bands in the oral ectoderm^{8,9}. In addition, *otx* has a conserved function in specifying anterior body regions and sense organs in insects and vertebrates, and in various other bilaterians²¹. We cloned the *Platynereis* *otx* homologue, which contains a homeobox with diagnostic amino acids and an N-terminal nonapeptide also found in chordates and lower Deuterostomia (Fig. 3a). In phylogenetic analysis, *Pd-otx* clearly clusters with other *otx* homologues (Fig. 3a; and data not shown). In early trochophora larva, *Pd-otx* is expressed strongly in cells in the oral region (peristomium), and in a few cells in the upper hemisphere (episphère) of the larva comprising developing apical neurons and sensory cells (22 h, Fig. 3b). *Pd-otx*-positive peristomial cells cover the stomodaeum, and laterally extend as two bands that surround the larva in an equatorial position (22 h, Fig. 3c). These cells demarcate the anterior and posterior margins of the peristomium, where two characteristic ciliary bands, the pre-oral prototroch and the postoral metatroch, form in trochophora larvae² (Fig. 1a; note that in the non-planktotrophic larva of *Platynereis*, metatrochal cilia appear only at post-larval stages). Expression remains associated with ciliary bands in the peristomium, but is not detected in another ciliary band (telotroch) in the posterior of the larval body. In the late trochophora, *Pd-otx* is expressed prominently in a transverse band of cells posteriorly abutting the stomodaeum that demarcates the posterior boundary of the peristomium (48 h, Fig. 3d). Furthermore, *Pd-otx* is expressed in a few cells of the stomodaeum itself, and it is still found in single cells in the developing episphère that represent subsets of apical sense organs and neurons. In the developing young worm, *Pd-otx* is still expressed in cells posteriorly abutting the prototroch in the peristomium (72 h, Fig. 3e) and in the prostomium—the 'tip' of the worm that develops from the larval episphère (data not shown). Notably, peristomium and prostomium together constitute the polychaete head². *Pd-otx* expression therefore outlines the developing polychaete head, with a sharp posterior boundary of expression (Fig. 3d). In this respect *Pd-otx* expression at postlarval stages resembles *otx* expression in anterior body regions in other bilaterians, and allows the affiliation of the polychaete head region with the *otx* territories of, for example, insects and vertebrates²¹. Taken together, *Pd-otx* expression in the trochophora larva is most prominent in the oral region, along pre- and postoral ciliary bands (red area in Fig. 1a). This equals *otx* expression along pre- and postoral ciliary bands in the oral ectoderm in the larvae of Deuterostomia^{8,9} (red area in Fig. 1b), in line with the evolutionary conservation of larval mouth regions. Of note, in mature larvae of Deuterostomia *otx* expression remains restricted to ciliary bands at the level of the mouth⁸. At these stages, pre- and postoral ciliary bands in deuterostome larvae form elaborate loops spanning the length of the larvae². *otx* expression, however, is absent from other parts of the ciliary loops in anterior and posterior body regions⁸.

To substantiate further the homology of larval foreguts in Protostomia and Deuterostomia, we cloned the *Platynereis* *goosecoid* homologue, which is a conserved marker for foregut roof (prechordal plate) in postgastrula vertebrate embryos¹⁰. *Platynereis* *goosecoid* (*Pd-gsc*) contains a conserved N-terminal heptapeptide (Fig. 4a) and, in phylogenetic analysis, clusters with *Drosophila* *goosecoid* within other *goosecoids* (data not shown). In *Platynereis*, *Pd-gsc* is expressed first in a small number of cells at the anterior margin of the blastopore that gives rise to the stomodaeum—the developing foregut. It is also expressed in two adjacent cells that later

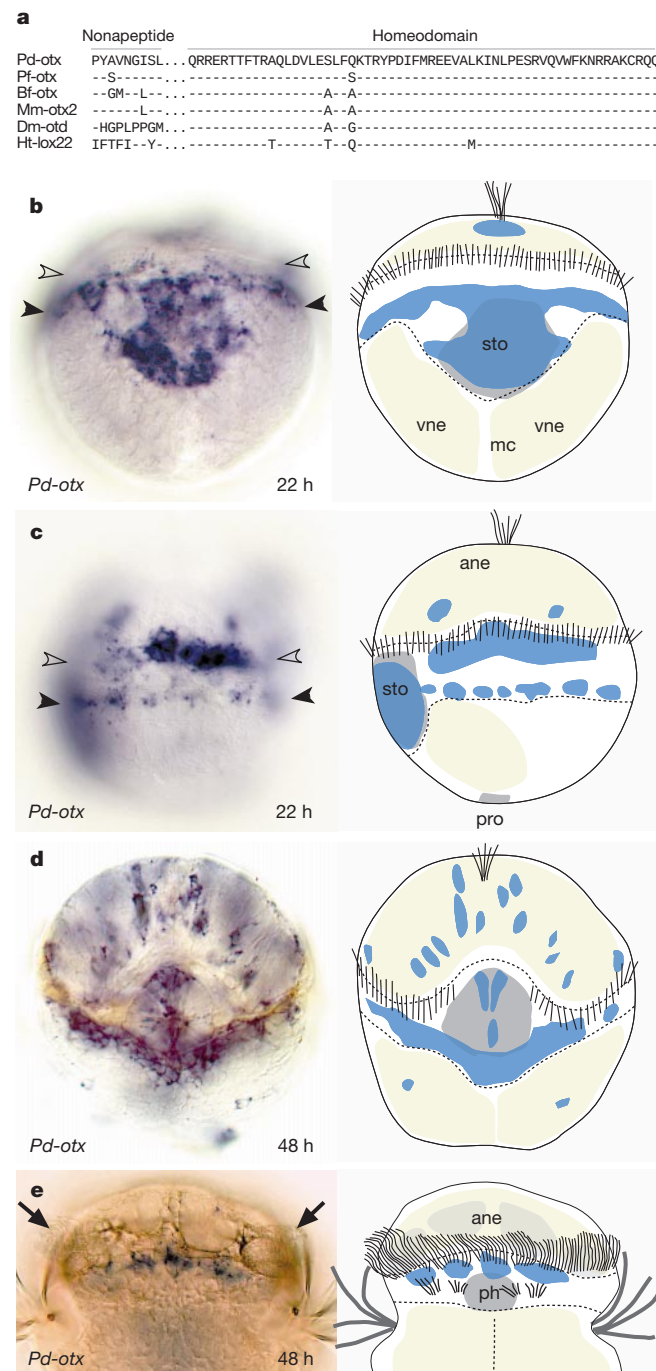


Figure 3 Expression of *Pd-otx*. **a**, Amino-acid sequence alignment of Otx nonapeptides and homeodomains. **b**, Ventral view of 22-h trochophora with *Pd-otx* expression (blue) in the stomodaeum and in more lateral cells (arrowheads) that form part of the mouth region (peristomium). **c**, Lateral view of 22-h trochophora showing expression in two equatorial bands of cells. Upper band of expression (open arrowheads) demarcates the anterior margin; lower band (closed arrowheads) demarcates the posterior margin of the peristomium. **d**, Ventral-anterior view of 48-h late trochophora with strong expression in the ventral peristomium. **e**, Ventral view of 72-h young worm. *Pd-otx* is still expressed along prototrochal cells bearing multiple cilia (arrows). Stippled lines outline the peristomium. Abbreviations and colours as in Fig. 2; Ht, *Helobdella triserialis*.

form part of the stomodaeal nervous system, from early to late trochophora stages (at 22 h, Fig. 4b, c; and 36 h, Fig. 4d). Expression of *goosecoid* in at least part of the developing foregut therefore occurs in a broad range of species in both Deuterostomia and Protostomia (the latter also including *Drosophila* with highly derived gastrulation^{22,23}).

From a comparative viewpoint, our data have far-reaching implications for the evolution of Bilateria. The shared, and very specific, expression of *brachyury* in ventral developing foreguts of the starfish bipinnaria⁵, echinoid pluteus¹⁴, enteropneust tornaria^{6,7} and polychaete trochophora larva suggests common ancestry (homology) of larval foreguts in Protostomia and Deuterostomia, despite the different developmental origin of these structures (Fig. 1). This hypothesis is supported by the common expression of *goosecoid* in the foreguts of various Bilateria. In addition, expression of *otx* along pre- and postoral ciliary bands in the mouth region of early and late holothurian bipinnaria⁸, enterop-

neust tornaria⁹ and polychaete trochophora larva also indicates homology, despite the differences in structure and function of larval ciliary bands in Protostomia and Deuterostomia². Although the alternative explanation of convergent evolution cannot be completely ruled out, it would require independent recruitment—at least twice—of *brachyury* for expression in the ventral portion of the stomodaeum, and of *otx* for expression in pre- and postoral ciliary bands in the oral ectoderm. This is unlikely given the specificity of the compared expression patterns. There is no *otx* expression along ciliary bands in anterior or posterior body regions of late bipinnaria⁸, nor in the posterior telotroch of trochophora larvae, ruling out a general requirement of *otx* for ciliary function. Most important, *brachyury* expression in the ventral portion of the stomodaeum and *otx* expression in ciliary bands occurs in primary, ciliary larvae only. In groups that lack these larval forms (such as tunicates²⁴, acraniates²⁵ and vertebrates²⁵ in Deuterostomia, and insects¹¹ and nematodes¹² in Protostomia), there are no *otx*-expressing ciliary bands, and likewise there is no stomodaeal *brachyury* expression. This observation establishes directionality in the evolutionary transition between indirect and direct development¹⁴: development through primary, ciliary larvae with tube-shaped guts and ciliary bands in the oral ectoderm seems to be ancestral for Protostomia, for Deuterostomia and thus for Bilateria as a whole. This idea was already propounded by Haeckel who speculated that primary, ciliary larvae in Bilateria should all belong, and trace back, to one single type—the so-called “vermilarva” with tube-shaped gut and circumoral ciliary bands²⁶. We propose that the difference in ontogeny of the larval tube-shaped guts in Protostomia versus Deuterostomia does not reflect a difference in their phylogeny. Tube-shaped guts in primary larvae evolved only once, and their ontogeny was subsequently modified in evolution. Our data do not indicate whether the ancestral form was protostome or deuterostome gut formation. Some exceptional cases of deuterostome gastrulation among molluscs and polychaetes however, suggest that the transition from protostomy towards deuterostomy is the probable pattern^{1,20}. Further, deuterostomy in vertebrates can be deduced from a gastrulation pattern with a slit-like blastopore²⁰.

Our work adds new characteristics to the Urbilateria, the stem species of all Bilateria¹⁹, which were proposed to show mesodermal segmentation²⁷, and to form an elaborated ventral nerve cord²⁸. We now propose that Urbilateria developed by means of a primary, ciliary larva with trochophora-type gastrulation forming a tube-shaped gut. *Brachyury* was expressed in the ventral portion of the stomodaeum, and *otx* along pre- and postoral ciliary bands. Possibly, this larva already possessed a population of *Brachyury*-expressing, ventral midline cells from which the notochord later originated in the chordate branch of evolution. □

Methods

Animal culture

We obtained larval stages from an established breeding culture at the University of Mainz⁴.

Isolation of *brachyury*, *otx* and *goosecoid* homologues

We isolated fragments of *Pd-bra*, *Pd-otx* and *Pd-gsc* from *Platynereis* by nested polymerase chain reaction after reverse transcription of embryonic messenger RNA (24 h). Degenerate primers for *Pd-bra*: forward, GNCGNATGTTTCNGTNC; nested forward, GGNYTNG AYAAYGCNATG; reverse, TTNGCRTCNAARRANGC; and nested reverse, GCNGTNACNGCDATRAAYTG. Degenerate primers for *Pd-otx*: forward, CAGMGMGGGARMGNACNACNTTYAC; reverse, ATSGAGGCGNGNSWCCADAT. Degenerate primers for *Pd-gsc*: forward, GCCGGIATGTTTACIATHGAYARAT; nested forward, CAYMGGACIATHTTYACIGAYGARCA; and reverse, TIGCICKICKRTTYTTR AACCA. These were used at 3 min 94 °C (1×), 1 min 94 °C, 1 min 42 °C, 1 min 72 °C (35×) and 10 min 72 °C (*Pd-bra*); at 1 min 94 °C, 2 min 42 °C, 4 min 72 °C (5×) and 1 min 94 °C, 2 min 47 °C, 4 min 72 °C (35×) and 10 min 72 °C (*Pd-otx*); or at 1 min 94 °C, 2 min 44 °C, 4 min 72 °C (5×) and 1 min 94 °C, 2 min 49 °C, 4 min 72 °C (35×) and 10 min 72 °C (*Pd-gsc*). Full length clones were obtained by 5' and 3' rapid amplification of cDNA ends on an embryonic complementary DNA library using sequence-specific primers (for *Pd-bra*: forward, ACAGTTTCCTGCTGGACCTGGTGAA; nested forward, GGCAATAAACTGTGTCTCCACGAAG; reverse, CGCTGGAAGTATGTGAACGGCGACT; and nested

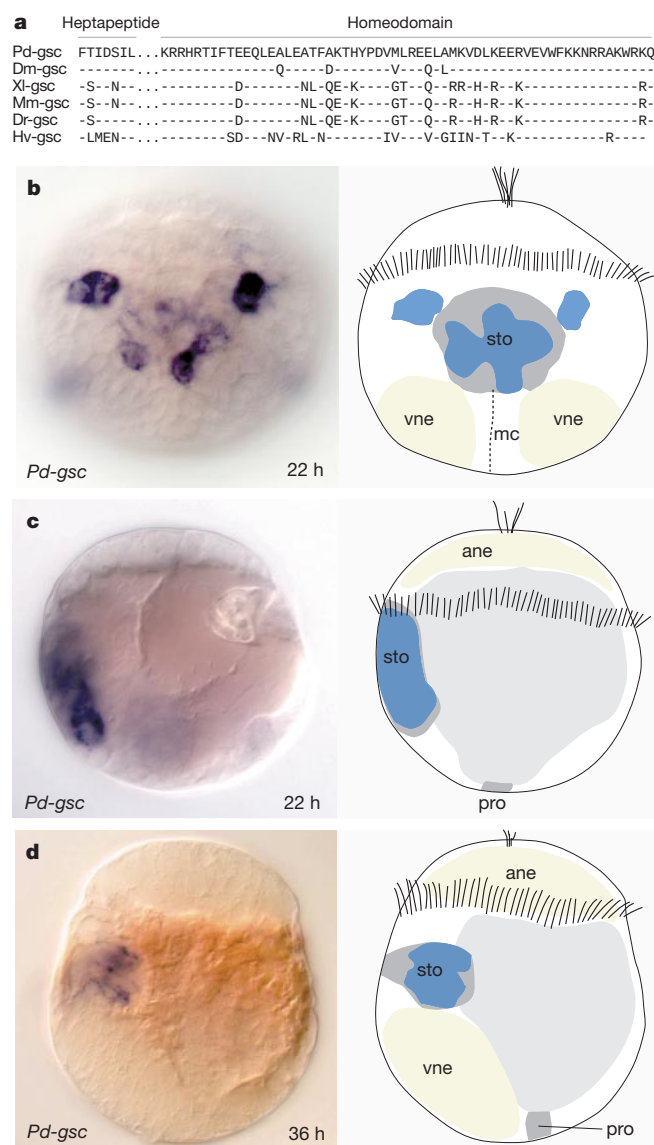


Figure 4 Expression of *Pd-gsc*. **a**, Amino-acid sequence alignment of Gsc heptapeptides and homeodomains. **b**, Ventral view of 22-h trochophora with expression in the stomodaeum. The two staining cells adjacent to the stomodaeum later form stomodaeal nervous system. **c**, Lateral view of 22-h trochophora. **d**, Lateral view of 36-h trochophora with *Pd-gsc* (blue) expression in the stomodaeum. Stippled line represents the ventral midline. Abbreviations and colours as in Fig. 2; *Xl*, *Xenopus laevis*, *Dr*, *Danio rerio*.

reverse, TGGGTGGAATCGTCTCTTCTCAT; for *Pd-otx*: forward, ACTTGCCAGAATCCAGAGTTCAAG; reverse, GGGACCATATGAGGCGAGTT; for *Pd-gsc*: forward, TGAACCAACCTCAACCCTCTTCTCT; reverse, GGACGATCTTTACGGATGAGCAA CTGG) and plasmid specific primers (T7 and T3) at 30 s 94 °C, 1 min 60 °C and 4 min 72 °C. We confirmed identity of the clones by sequencing (EMBL Nucleotide Sequence Database accession numbers: *Pd-bra*, PDU289022; *Pd-otx*, AJ278856; and *Pd-gsc*, PDU289023).

Phylogenetic analysis

We obtained protein sequences of a selected number of species from the database and aligned them using CLUSTALW. We used these alignments to calculate a phylogenetic tree using the maximum likelihood program PUZZLE²⁹.

Whole-mount *in situ* analysis

We fixed embryos in 4% paraformaldehyde /2 × PBS-Tween for 1–4 h. An established *in situ* hybridization protocol³⁰ was followed with the modification of Proteinase K treatment in 100 µg ml⁻¹ for 4 min (24 h larvae), or 10 min (72 h young worms). After staining, embryos were refixed in paraformaldehyde /2 × PBS-Tween, washed and cleared in 80% glycerol. We mounted embryos in glycerol and took pictures under Nomarsky optics using a Zeiss Axiophot.

Received 21 August; accepted 12 October 2000.

- Willmer, P. *Invertebrate Relationships* (Cambridge Univ. Press, Cambridge, 1990).
- Nielsen, C. *Animal Evolution: Interrelationships of the Living Phyla* (Oxford Univ. Press, Oxford, 1995).
- Grobbs, K. Die systematische Einteilung des Tierreichs. *Verh. Zool. Bot. Ges. Wien* **58**, 491–511 (1908).
- Dorrestein, A. W. C., O'Grady, B., Fischer, A., Porchet-Henere, E. & Boilly-Marer, Y. Molecular specification of cell lines in the embryo of *Platynereis* (Annelida). *Roux's Arch. Dev. Biol.* **202**, 264–273 (1993).
- Shoguchi, E., Satoh, N. & Maruyama, Y. K. Pattern of *Brachyury* gene expression in starfish embryos resembles that of hemichordate embryos but not of sea urchin embryos. *Mech. Dev.* **82**, 185–189 (1999).
- Peterson, K. J., Cameron, R. A., Tagawa, K., Satoh, N. & Davidson, E. H. A comparative molecular approach to mesodermal patterning in basal deuterostomes: the expression pattern of *Brachyury* in the enteropneust hemichordate *Ptychodera flava*. *Development* **126**, 85–95 (1999).
- Tagawa, K., Humphreys, T. & Satoh, N. Novel pattern of *Brachyury* gene expression in hemichordate embryos. *Mech. Dev.* **75**, 139–143 (1998).
- Shoguchi, E., Harada, Y., Numakunai, T. & Satoh, N. Expression of the *Otx* gene in the ciliary bands during sea cucumber embryogenesis. *Genesis* **27**, 58–63 (2000).
- Harada, Y. *et al.* Developmental expression of the hemichordate *otx* ortholog. *Mech. Dev.* **91**, 337–339 (2000).
- Steinbeisser, H. & De Robertis, E. M. *Xenopus goosecoid*: a gene expressed in the prechordal plate that has dorsalizing activity. *C. R. Acad. Sci.* **316**, 959–971 (1993).
- Kispert, A., Herrmann, B. G., Leptin, M. & Reuter, R. Homologs of the mouse *Brachyury* gene are involved in the specification of posterior terminal structures in *Drosophila*, *Tribolium*, and *Locusta*. *Genes Dev.* **8**, 2137–2150 (1994).
- Woolard, A. & Hodgkin, J. The *Caenorhabditis elegans* fate-determining gene *mab-9* encodes a T-box protein required to pattern the posterior hindgut. *Genes Dev.* **14**, 596–603 (2000).
- Kispert, A. & Herrmann, B. G. The *Brachyury* gene encodes a novel DNA binding protein. *EMBO J.* **12**, 3211–3220 (1993).
- Peterson, K. J., Cameron, R. A. & Davidson, E. H. Bilateral origins: significance of new experimental observations. *Dev. Biol.* **219**, 1–17 (2000).
- Kusch, T. & Reuter, R. Functions for *Drosophila brachyenteron* and *forkhead* in mesoderm specification and cell signalling. *Development* **126**, 3991–4003 (1999).
- Smith, J. *Brachyury* and the T-box genes. *Curr. Opin. Genet. Dev.* **7**, 474–480 (1997).
- Wilson, E. B. The cell-lineage of *Nereis*. A contribution to the cytogeny of the annelid body. *J. Morphol.* **6**, 361–480 (1892).
- Wu, L. H. & Lengyel, J. A. Role of caudal in hindgut specification and gastrulation suggests homology between *Drosophila* amnioproctodeal invagination and vertebrate blastopore. *Development* **125**, 2433–2442 (1998).
- De Robertis, E. M. & Sasai, Y. A common plan for dorsoventral patterning in Bilateria. *Nature* **380**, 37–40 (1996).
- Arendt, D. & Nübler-Jung, K. Dorsal or ventral: similarities in fate maps and gastrulation patterns in annelids, arthropods and chordates. *Mech. Dev.* **61**, 7–21 (1997).
- Hirth, F. & Reichert, H. Conserved genetic programs in insect and mammalian brain development. *BioEssays* **21**, 677–684 (1999).
- Hahn, M. & Jäckle, H. *Drosophila goosecoid* participates in neural development but not in body axis formation. *EMBO J.* **15**, 3077–3084 (1996).
- Goriely, A. *et al.* A functional homologue of *goosecoid* in *Drosophila*. *Development* **122**, 1641–1650 (1996).
- Yasuo, H. & Satoh, N. Conservation of the developmental role of *Brachyury* in notochord formation in a urochordate, the ascidian *Halocynthia roretzi*. *Dev. Biol.* **200**, 158–170 (1998).
- Holland, P. W., Koschorz, B., Holland, L. Z. & Herrmann, B. G. Conservation of *Brachyury* (T) genes in amphioxus and vertebrates: developmental and evolutionary implications. *Development* **121**, 4283–4291 (1995).
- Haeckel, E. in *Systematische Phylogenie. 2. Teil: Systematische Phylogenie der wirbellosen Tiere (Invertebrata)*. 259–347 (Georg Reimer, Berlin, 1896).
- De Robertis, E. M. Evolutionary biology. The ancestry of segmentation. *Nature* **387**, 25–26 (1997).
- Arendt, D. & Nübler-Jung, K. Comparison of early nerve cord development in insects and vertebrates. *Development* **126**, 2309–2325 (1999).
- Strimmer, K. & Von Haeseler, A. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl Acad. Sci. USA* **94**, 6815–6819 (1997).
- Loosli, F., Köster, R. W., Carl, M., Krone, A. & Wittbrodt, J. Six3, a medaka homologue of the

Drosophila homeobox gene *sine oculis* is expressed in the anterior embryonic shield and the developing eye. *Mech. Dev.* **74**, 159–164 (1998).

Acknowledgements

We thank A. A. W. Dorrestein, F. Loosli and R. Rieger for discussions; S. Cohen, B. Hobmayer, T. Holstein, T.-E. Rusten and L. Teixeira for comments on the manuscript; and members of the Wittbrodt laboratory for support. cDNA libraries were provided by C. Heimann, University of Mainz. This work was supported by a fellowship from the European Molecular Biology Organisation (EMBO) (D.A.), and by grants from the Deutsche Forschungsgemeinschaft (DFG) Schwerpunkt "Evolution entwicklungsbiologischer Prozesse" (U.T. and J.W.).

Correspondence and requests for materials should be addressed to J.W. (e-mail: Jochen.Wittbrodt@EMBL-Heidelberg.de) or U.T. (e-mail: technau@bio.tu-darmstadt.de).

Self-motion and the perception of stationary objects

Mark Wexler, Francesco Panerai, Ivan Lamouret & Jacques Droulez

Laboratoire de Physiologie de la Perception et de l'Action, Collège de France, 11 place Marcelin Berthelot, 75005 Paris, France

One of the ways that we perceive shape is through seeing motion^{1–3}. Visual motion may be actively generated (for example, in locomotion), or passively observed. In the study of the perception of three-dimensional structure from motion, the non-moving, passive observer in an environment of moving rigid objects has been used as a substitute¹ for an active observer moving in an environment of stationary objects; this 'rigidity hypothesis' has played a central role in computational and experimental studies of structure from motion^{4,5}. Here we show that this is not an adequate substitution because active and passive observers can perceive three-dimensional structure differently, despite experiencing the same visual stimulus: active observers' perception of three-dimensional structure depends on extraretinal information about their own movements. The visual system thus treats objects that are stationary (in an allocentric, earth-fixed reference frame) differently from objects that are merely rigid. These results show that action makes an important contribution to depth perception, and argue for a revision of the rigidity hypothesis to incorporate the special case of stationary objects.

The original work comparing actively produced to passively observed motion parallax^{6,7} found structure from motion (SFM) performance that depended on retinal information alone: non-moving observers receiving optical information similar to that received by active observers had similar response thresholds. Other studies have found that self-motion helps to resolve discrete symmetries in optic flow^{8,9}, or to decrease integration times in SFM¹⁰.

In the first experiment we tested extraretinal contributions to the extraction of depth from motion by means of a cue–conflict paradigm, in which motion parallax cues to three-dimensional (3D) structure were weighed against conflicting linear perspective (that is, the assumption that lines nearly parallel or perpendicular in the image are actually parallel or perpendicular in 3D space). The observer saw a planar 3D grid in motion, and estimated its tilt (the direction of its projected normal in the frontoparallel plane). Motion parallax could be actively produced or passively observed. In the active case, parallax was due to the observer's head movements around a virtual object; in the passive case, the observer remained still while watching a replay of the optic flow from a preceding active trial. The tilt of the plane defined by perspective

cues differed from the tilt defined by motion cues by $\Delta T = 0^\circ, 45^\circ, 90^\circ, 135^\circ$ or 180° . further details and examples of stimuli are shown in Fig. 1.

In their tilt judgements, subjects could ignore one cue, switch between cues, or base their responses on a weighted average of the two cues—although theoretical considerations suggest that for large conflicts such as ours, cue averaging would not be optimal¹¹.

If on a particular trial the observer relied on the motion cue alone, he or she would perceive a spatially irregular structure that contradicted laws of linear perspective¹², but one that underwent rigid 3D motion. If the observer relied on the perspective cue alone, he or she would perceive a structure that was more spatially regular, but that underwent deformation in time, thus violating the rigidity assumption that is supposed to underlie SFM^{5,13,14} (but see also

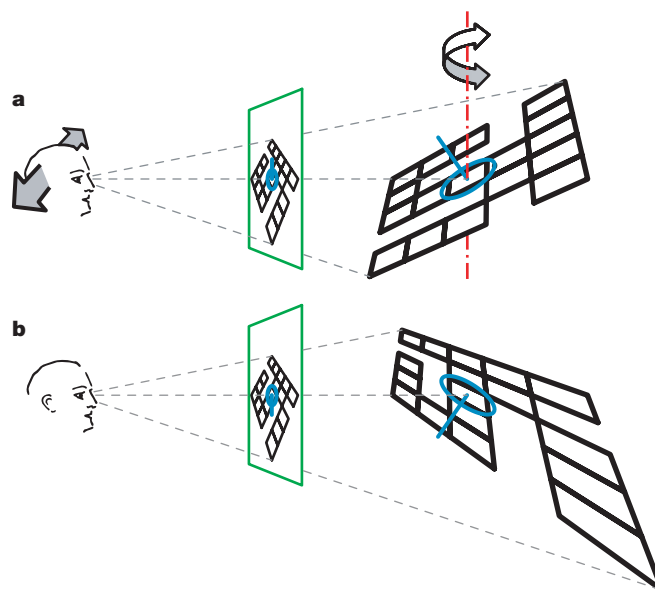


Figure 1 Stimuli used in the experiments. **a**, A no-conflict stimulus, where both perspective and motion cues indicate a surface tilted upwards. Motion was due either to subject's head movements, or to object rotation. Both the virtual object and its projection are shown. The subject's task was to line up a probe (shown in blue) with the surface. **b**, A conflict stimulus, generated by back-projecting the no-conflict stimulus onto a

different plane. Although the virtual object is now tilted downwards, in its central position the projection is identical to the no-conflict stimulus; thus, perspective cues indicate upward tilt, while motion cues indicate downward tilt. In this case, the tilt conflict ΔT would be 180° .

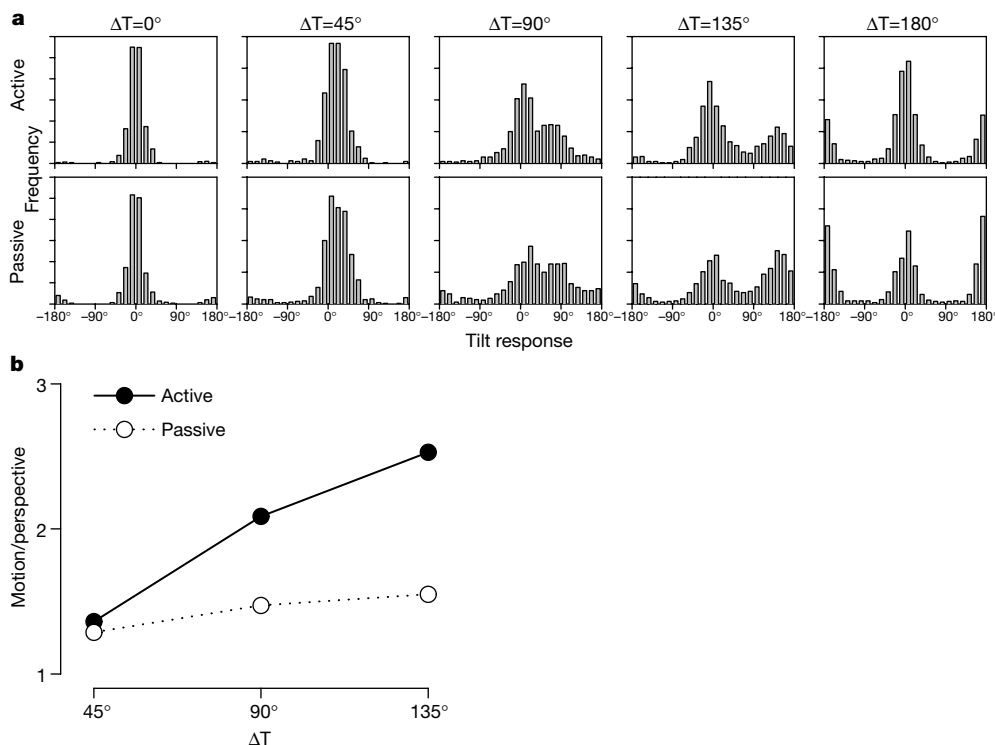


Figure 2 Results of experiment 1, averaged over subjects. **a**, Frequencies of tilt responses, in the active and passive conditions, for no tilt conflict ($\Delta T = 0^\circ$) and tilt conflicts $\Delta T = 45^\circ, 90^\circ, 135^\circ$ and 180° . Responses are adjusted so that motion tilt is

always at 0° , and so that perspective tilt is positive. **b**, The ratio of motion- to perspective-based responses in the active and passive conditions, as a function of tilt conflict.

refs 15–18)—as in the well-known Ames window phenomenon^{19–23}.

Tilt responses in experiment 1 are shown in Fig. 2a. If the response is in accordance with motion cues, we would expect a peak at 0° (all angles are defined with respect to motion tilt); if the response is in accordance with perspective cues, the peak would be at ΔT . Furthermore, motion cues can yield an ‘inverted’ response at $\pm 180^\circ$, owing to the approximate symmetry of optic flow under the simultaneous inversion of motion and tilt; this symmetry is exact in the case of parallel projection but, in our case, the inverted solution is not perfectly rigid^{8,9}. As can be seen from Fig. 2a, the responses are based on a multimodal mixture of perspective and motion cues (the multimodality is present in individual subject data), and, as predicted for large conflicts¹¹, not on cue averages.

The main effect of self-motion is on the relative strength of the motion and perspective cues: responses in accordance with motion cues are more frequent in the active than in the passive condition. In order to quantify this self-motion effect, we counted trials with responses based on motion cues, and those with responses based on perspective cues (see Fig. 2b). Motion responses are defined as those falling within $\pm 22.5^\circ$ bins of 0° and 180°, perspective responses as those around ΔT . (Conditions $\Delta T = 0^\circ$ and 180° , where there is a partial confound between the two types of responses, are excluded.) A three-way analysis of variance on ΔT , cue and self-motion variables showed a significant cue $\times$ self-motion interaction ($F_{1,7} = 6.53$, $P < 0.05$). A similar effect of self-motion was predicted²⁴ in discussing the Ames window (see ref. 25 for related work).

The other effect of self-motion is fewer inverted responses in the active condition: for $0^\circ \leq \Delta T \leq 135^\circ$, tilt responses in 7.1% of the trials are inverted (that is, they cluster around $T = 180^\circ$) in the active condition, whereas 20.3% are inverted in the passive condition ($F_{1,7} = 25.4$, $P < 0.01$)^{8,9}. On the other hand, the precision of tilt judgements was no different in the active than in the passive case. This can be shown by fitting the distributions in Fig. 2a with sums of gaussians centred about $T = 0$ and $T = \Delta T$. Mean widths of the $T = 0$ peaks were 27.3° and 27.1° for the active and passive cases, respectively, with the difference not significant. The similar precision of SFM in active and passive conditions echoes previous results^{6,9}, which also found that perception of 3D structure was not more precise in active than in passive observers.

A crucial difference between what subjects perceive in the active and passive conditions are spatial attributes of the object in an allocentric reference frame. When the observer utilizes motion depth cues at the expense of conflicting perspective, he or she perceives a rigid 3D object. In the active case, this rigid object is also stationary in an allocentric, earth-fixed reference frame, whereas in the passive case, the object specified by motion cues is no less rigid but undergoes movement in space^{26,27}. In principle, stationarity—as opposed to rigidity—is impossible to reliably determine from optic information alone. (Physiologically, the visual system may sometimes be fooled into judging an object to be stationary from large-angle optic flow, as in the case ofvection.) We propose that motion cues are enhanced in the active condition over the passive condition because in the former case they lead to percepts that are not only rigid, but also stationary: the ‘stationarity hypothesis’. The relative dearth of inverted—and therefore non-stationary—solutions in active trials in experiment 1 is further, indirect evidence for the stationarity hypothesis. On the other hand, motion cues could simply be enhanced for an observer in motion. There is no way to distinguish the stationarity and the ‘motion-enhancement’ hypotheses in experiment 1, where moving observers always perceive stationary objects from motion cues.

We directly tested the stationarity hypothesis in a second experiment, which used the same active/passive cue–conflict paradigm as experiment 1. But here, on some active trials (‘twist’ trials), the virtual object was not stationary but underwent oscillations about a

horizontal axis. In the absence of cue conflict, in the twist trials active observers correctly perceived an object undergoing rigid oscillatory motion, in synchrony with their own head movements. The stationarity hypothesis predicts that in active, non-stationary trials the use of motion cues would be reduced relative to active, stationary trials. On the other hand, if the enhancement of motion cues in the active condition of experiment 1 were due to mere presence of self-motion, we would expect no effect of non-stationarity.

As can be seen from Fig. 3a, responses on no-twist trials are similar to those in experiment 1. In twist trials without cue conflict, there is still a sharp peak about $\Delta T = 0$, showing that subjects were able to do the task even in the presence of a non-stationary stimulus synchronized with their own motion. In twist trials with cue conflict, on the other hand, the dominance of motion cues seen in the stationary case disappears, as predicted by the stationarity hypothesis, and is replaced by approximately equal peaks around motion and perspective cues, reminiscent of the passive condition in experiment 1. To quantify this result, we have counted motion- and perspective-based trials, as in experiment 1; their ratios are shown in Fig. 3b. The effect of stationarity on the active case was significant: in the $\Delta T = 90^\circ$ case, there was an interaction between cue and twist variables ($F_{1,7} = 44.0$, $P < 0.001$). This effect is not due simply to the change in the axis of rotation that is introduced by the twist, as seen from a cue $\times$ twist $\times$ self-motion interaction ($F_{1,7} = 18.0$, $P < 0.01$). As predicted by the stationarity hypothesis, the highest ratio was in the one case (active, no-twist) where motion cues yielded a spatially stationary object.

In two experiments we show two results. The more general result

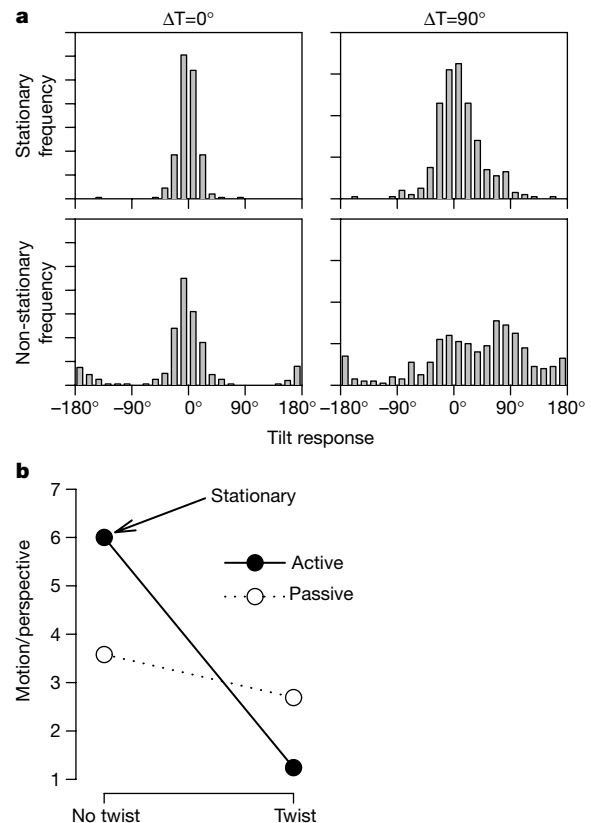


Figure 3 Results of experiment 2, averaged over subjects. **a**, Frequencies of tilt errors in active stationary and active non-stationary conditions, with and without tilt conflict. Passive results were similar to those in experiment 1, showing no effect of twist. **b**, Ratio of motion- to perspective-based responses with and without twist in active and passive conditions, with tilt conflict. Responses are defined as motion-based when tilt falls within $\pm 45^\circ$ bins about 0° and 180° and as perspective-based when tilt falls within $\pm 45^\circ$ of 90°.

is that extraretinal, self-motion information is incorporated into visual judgements of three-dimensional structure. The visual stimulation in the active and passive conditions of experiment 1 is the same, yet the active observer responds more frequently on the basis of motion cues than does the passive observer. Thus the effect of self-motion on spatial vision does not reduce to its modification of optic flow. The second result builds on the first: the relative enhancement of motion cues only occurs when they indicate objects that are stationary in an allocentric reference frame. It seems that the visual system is biased towards perceiving stationary objects, even when their image deforms as a result of observer motion. Physiological findings point to the existence of allocentric coding in mammalian brains²⁸ that could be involved in this process. Our results would point to a revision of the rigidity concept in SFM, and, more generally, the inadequacy of excluding observer motion in the analysis of spatial vision. □

Methods

Experiments were performed in monocular conditions in darkness. Translations of the subject's dominant eye were measured by a head tracker²⁹, sampled on active trials by a personal computer that displayed a virtual object, polar-projected for the current eye position. Data were also stored for use in subsequent passive trials.

In active trials, subjects performed lateral oscillatory head movements about a central point, whose perpendicular distance to the monitor was between 35 and 45 cm. In each trial three cycles were performed, with the stimulus appearing after the first half-cycle. Mean maximum displacement amplitudes about stimulus centre were $19.7 \pm 7.1^\circ$ and $17.7 \pm 4.3^\circ$, and mean periods were 1.9 ± 0.5 and 2.0 ± 0.5 s in experiments 1 and 2, respectively.

Stimuli were polar projections of a virtual 3D object, a partial grid of at most 10×10 square cells, each 1 cm in length, with a slant of 45° and a randomly chosen tilt. A certain number of distinct cells were randomly chosen and removed from the grid; stimuli consisted of 5, 10, 30, 60 or 100 cells in experiment 1, and of 10 cells in experiment 2. The centre of the grid was at the point on the monitor opposite the subject's initial eye position. Stimuli were drawn as one-pixel-thick white lines on a black background, with a small red fixation point at the centre.

In order to generate conflict stimuli, the initial grids, which had tilt T_p , underwent a projection from the subject's initial eye position onto a plane passing through the grid centre, with a slant of 45° but with a tilt T_m . The resulting virtual object was an irregular grid (similar to the 'Ames window' or 'Ames chair'). (In the analysis of the results, response tilts are measured with respect to T_m , with sign defined so that so that $\Delta T = T_p - T_m \geq 0$.)

Blocks of passive trials were alternated with blocks of active trials. In passive trials, subjects experienced similar optic flow as in active trials, but without head movement. Each passive trial corresponded to a preceding active trial, in that the virtual object used to generate the stimuli was identical, and that rotations of the virtual object about its centre with respect to the subject's eye were identical in passive and active trials. Let the initial eye position be $\mathbf{r}_0$ and the eye position at a given moment of an active trial be $\mathbf{r}$ (relative to the centre of the virtual object used to generate the stimulus); let $\theta = \arccos[(\mathbf{r}_0 \cdot \mathbf{r})/(|\mathbf{r}_0||\mathbf{r}|)]$ be the angle between $\mathbf{r}_0$ and $\mathbf{r}$, and let $\mathbf{a}$ be the axis generating the rotation from $\mathbf{r}_0$ to $\mathbf{r}$ (that is, $\mathbf{a}$ is parallel to $\mathbf{r}_0 \times \mathbf{r}$). At the corresponding moment during the passive trial, the virtual object was rotated by angle $-\theta$ about axis $\mathbf{a}$ before being projected.

In 'twist' trials in experiment 2, the retinal optical flow underwent a 90° rotation, generated as follows. The virtual object underwent the same rotation as the eye about its centre (that is, by angle θ about axis $\mathbf{a}$). Then $\mathbf{r}_0$ was rotated by the twist angle about $\mathbf{r}$ to yield $\mathbf{r}'_0$, and a new axis $\mathbf{a}' = \mathbf{r}'_0 \times \mathbf{r}$ calculated. Finally, the virtual object was rotated by angle $-\theta$ about the new axis $\mathbf{a}'$ and projected. Thus, if the twist angle were zero, the object would remain stationary (that is, in an egocentric frame it rotated about an approximately vertical axis, as in experiment 1), whereas for a 90° twist the object underwent the same rotations in the egocentric frame, but about an approximately horizontal axis.

After the disappearance of the stimulus, the projection of a virtual probe object was displayed on the monitor. The subjects' task was to align the probe—which was comprised of a circle and a perpendicular line—with the perceived plane, using a joystick.

In experiment 1, five active and five passive blocks were performed, with 100 trials in each block. In experiment 2, one active and one passive block were performed, with 160 trials per block. Eight naive volunteers participated as subjects in each of experiments 1 and 2.

Received 6 September; accepted 23 October 2000.

- Wallach, H. & O'Connell, D. N. The kinetic depth effect. *J. Exp. Psychol.* **45**, 205–217 (1953).
- Braunstein, M. L. Depth perception in rotating dot patterns. *J. Exp. Psychol. Hum. Percept. Perf.* **72**, 415–420 (1962).
- Johansson, G. Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* **14**, 210–211 (1973).
- Ullman, S. *The Interpretation of Visual Motion* (MIT Press, Cambridge, 1979).
- Koenderink, J. J. Optic flow. *Vision Res.* **26**, 161–179 (1986).
- Rogers, B. & Graham, M. Motion parallax as an independent cue for depth perception. *Perception* **8**, 125–134 (1979).
- Rogers, B. & Graham, M. Similarities between motion parallax and stereopsis in human depth perception. *Vision Res.* **22**, 261–270 (1982).

- Rogers, S. & Rogers, B. J. Visual and nonvisual information disambiguate surfaces specified by motion parallax. *Percept. Psychophys.* **52**, 446–452 (1992).
- Dijkstra, T. M., Cornilleau-Péres, V., Gielen, C. C. & Droulez, J. Perception of three-dimensional shape from ego- and object-motion: comparison between small- and large-field stimuli. *Vision Res.* **35**, 453–462 (1995).
- van Damme, W. J. & van de Grind, W. A. Non-visual information in structure-from-motion. *Vision Res.* **36**, 3119–3127 (1996).
- Landy, M. S., Maloney, L. T., Johnston, E. B. & Young, M. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res.* **35**, 389–412 (1995).
- Attneave, F. & Frost, R. The determination of perceived tridimensional orientation by minimum criteria. *Percept. Psychophys.* **6**, 391–396 (1969).
- Longuet-Higgins, H. C. & Prazdny, K. The interpretation of a moving retinal image. *Proc. R. Soc. Lond. B* **208**, 385–397 (1980).
- Todd, J. T. Visual information about rigid and non-rigid motion: a geometric analysis. *J. Exp. Psychol. Hum. Percept. Perf.* **8**, 238–252 (1982).
- Wallach, H., Weisz, A. & Adams, P. A. Circles and derived figures in rotation. *Am. J. Psychol.* **69**, 48–59 (1956).
- Adelson, E. H. Rigid objects that appear highly non-rigid. *Invest. Ophthalm. Visual Sci.* **26** (Suppl.), (1985).
- Sinha, P. & Poggio, T. Role of learning in three-dimensional form perception. *Nature* **384**, 460–463 (1996).
- Sparrow, J. E. & Stine, W. W. The perceived rigidity of rotating eight-vertex geometric forms: extracting nonrigid structure from rigid motion. *Vision Res.* **38**, 541–556 (1998).
- Ames, A. Visual perception and the rotating trapezoidal window. *Psychol. Monogr.* **65**, (1951).
- Ittelson, W. H. *The Ames Demonstrations in Perception* (Princeton Univ. Press, Princeton, 1952).
- O'Brien, J. & Johnston, A. When texture takes precedence over motion in depth perception. *Perception* **29**, 437–452 (2000).
- Wade, N. J. & Hughes, P. Fooling the eyes: trompe l'oeil and reverse perspective. *Perception* **28**, 1115–1119 (1999).
- Papathomas, T. V. See how they turn: False depth and motion in Hughes's reverspectives. *Proc. SPIE* **3959**, 506–517 (2000).
- Gibson, J. J. *The Ecological Approach to Visual Perception* (Houghton-Mifflin, Boston, 1979).
- Reinhardt-Rutland, A. H. Perceiving surface orientation: Pictorial information based on rectangularity can be overridden during observer motion. *Perception* **22**, 335–341 (1993).
- Wallach, H. Perceiving a stable environment when one moves. *Annu. Rev. Psychol.* **38**, 1–27 (1987).
- Ono, H. & Steinbach, M. J. Monocular stereopsis with and without head movement. *Percept. Psychophys.* **48**, 179–187 (1990).
- Snyder, L. H., Grieve, K. L., Brotschie, P. & Andersen, R. A. Separate body- and world-referenced representations of visual space in parietal cortex. *Nature* **394**, 887–891 (1998).
- Panerai, E., Hanneton, S., Droulez, J. & Cornilleau-Péres, V. A 6-dof device to measure head movements in active vision experiments: Geometric modeling and metric accuracy. *J. Neurosci. Meth.* **90**, 97–106 (1999).

Acknowledgements

We thank M. Ehrette and P. Leboucher for designing and building the head tracker.

Correspondence and requests for materials should be addressed to M.W. (e-mail: wexler@ccr.jussieu.fr).

Adaptive regulation of neuronal excitability by a voltage-independent potassium conductance

Stephen G. Brickley*, Victoria Revilla†, Stuart G. Cull-Candy*, William Wisden† & Mark Farrant*

* Department of Pharmacology, University College London, London WC1E 6BT, UK

† Medical Research Council Laboratory of Molecular Biology, Medical Research Council Centre, Cambridge CB2 2QH, UK

Many neurons receive a continuous, or 'tonic', synaptic input, which increases their membrane conductance, and so modifies the spatial and temporal integration of excitatory signals^{1–3}. In cerebellar granule cells, although the frequency of inhibitory synaptic currents is relatively low, the spillover of synaptically released GABA (γ -aminobutyric acid)⁴ gives rise to a persistent conductance mediated by the GABA_A receptor^{5–7} that also modifies the excitability of granule cells⁸. Here we show that this tonic conductance is absent in granule cells that lack the $\alpha 6$ and δ -subunits

of the GABA_A receptor. The response of these granule cells to excitatory synaptic input remains unaltered, owing to an increase in a 'leak' conductance, which is present at rest, with properties characteristic of the two-pore-domain K⁺ channel TASK-1 (refs 9–12). Our results highlight the importance of tonic inhibition mediated by GABA_A receptors, loss of which triggers a form of homeostatic plasticity leading to a change in the magnitude of a voltage-independent K⁺ conductance that maintains normal neuronal behaviour.

GABA_A-receptor-mediated inhibition of granule cells (GCs) is critical for normal cerebellar function^{13,14} and motor coordination¹⁵. Two forms of GABA_A receptor activation are seen during GC development: phasic inhibitory postsynaptic currents (IPSCs) and a spillover-mediated component^{4–7} that generates tonic shunting inhibition⁸. Although GCs express several GABA_A receptor types, an extrasynaptic receptor population of high affinity¹⁶, containing both $\alpha 6$ and δ -subunits, is progressively expressed during GC development^{17,18}. To establish whether these receptors underlie the emergence of the tonic conductance, we made recordings from GCs in brain slices prepared from mice in which the $\alpha 6$ gene was disrupted ($\alpha 6^{-/-}$)¹⁹. Because of their preferential assembly, this deletion causes a complete loss of both the $\alpha 6$ and the δ -subunit from the GC surface membrane^{19,20}, but other subunits ($\alpha 1$, $\beta 2$, $\beta 3$, $\gamma 2$) persist at synaptic locations²⁰. GCs exhibited a low frequency (~ 0.2 Hz) of both spontaneous IPSCs and excitatory synaptic currents (EPSCs). In wild-type mice, the selective GABA_A receptor antagonist SR95531 abolished IPSCs, and produced an outward current ($n = 32$; Fig. 1a, c, d), reflecting the block of tonically active receptors^{5–8}. The magnitude of the conductance blocked by

SR95531 (G_{GABA}) increased with age⁸. In contrast, GCs from $\alpha 6^{-/-}$ mice displayed a normal frequency of IPSCs, but G_{GABA} was absent ($n = 24$; Fig. 1b, c, d). This result shows that the activation of GABA_A receptors containing the $\alpha 6$ subunit ($\alpha 6\beta\gamma$ and/or $\alpha 6\beta\delta$ receptors^{19,20}) generates G_{GABA} .

Paradoxically, $\alpha 6^{-/-}$ mice exhibit normal motor function^{19,21,22}, calling into question the significance of G_{GABA} in the control of GC excitability. In voltage recordings from P35 wild-type GCs, however, SR95531 caused a significant reduction in the current required to reach the action-potential threshold ($-24 \pm 6\%$, $n = 8$; $P < 0.05$, Fig. 2a, b) and an increase in the amplitude of spontaneous excitatory synaptic potentials (EPSPs; $+16 \pm 3\%$; $n = 5$, $P < 0.01$, Fig. 2e), consistent with the observed decrease in input conductance ($-18 \pm 3\%$, $n = 15$; $P < 0.001$, Wilcoxon matched pair). As expected, given the absence of G_{GABA} , $\alpha 6^{-/-}$ GCs were not affected by SR95531 (threshold current $-6 \pm 7\%$, $n = 8$; input conductance $-2 \pm 3\%$, $n = 14$; EPSP amplitude $+3 \pm 10\%$, $n = 3$, Fig. 2c, d and f). These data show that tonic activation of $\alpha 6$ -containing GABA_A receptors is capable of altering the response of GCs to excitatory synaptic input, confirming that G_{GABA} modulates GC excitability⁸.

Despite the loss of G_{GABA} , $\alpha 6^{-/-}$ GCs did not differ from wild-type GCs in their membrane properties, responses to depolarizing current injection or EPSP properties (Table 1). The fact that GC excitability was indistinguishable suggests that changes in one or more conductance must compensate for the loss of G_{GABA} . We did

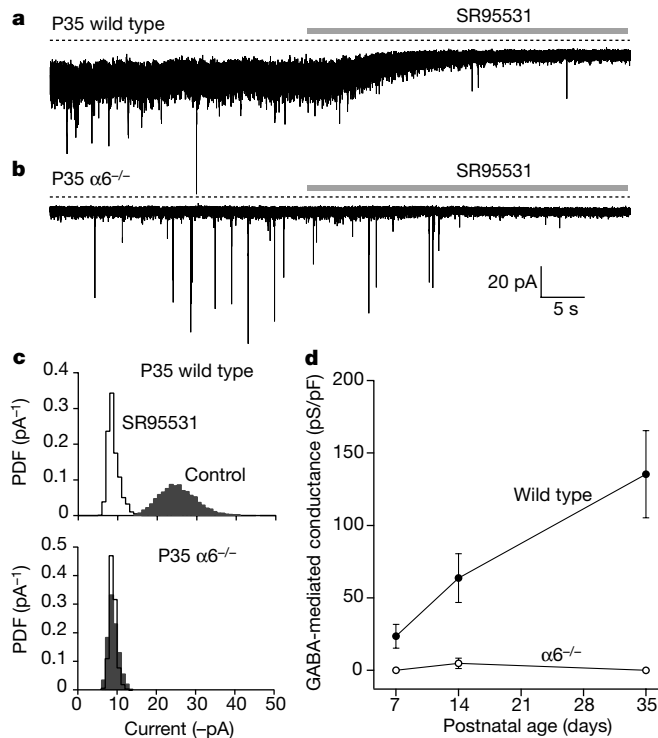


Figure 1 Absence of G_{GABA} in $\alpha 6^{-/-}$ GCs. **a**, Current record from a P35 wild-type GC (-70 mV) during application of $10 \mu\text{M}$ SR95531 (bar). The dashed line indicates the zero current level. **b**, Same as **a**, but from an $\alpha 6^{-/-}$ GC. In both cases, IPSCs (downward deflections) are abolished, but note the difference in the effect of SR95531 on the holding current. Synaptic currents remaining in SR95531 are EPSCs (rapid kinetics and blocked by CNQX). **c**, Probability density functions (PDF) of holding current in a wild-type and an $\alpha 6^{-/-}$ GC. The control and SR95531 peak currents were used to calculate G_{GABA} . **d**, Developmental change in G_{GABA} for wild-type and $\alpha 6^{-/-}$ GCs. Symbols indicate mean conductance ($n = 4$ –20 cells) and error bars indicate s.e.m.

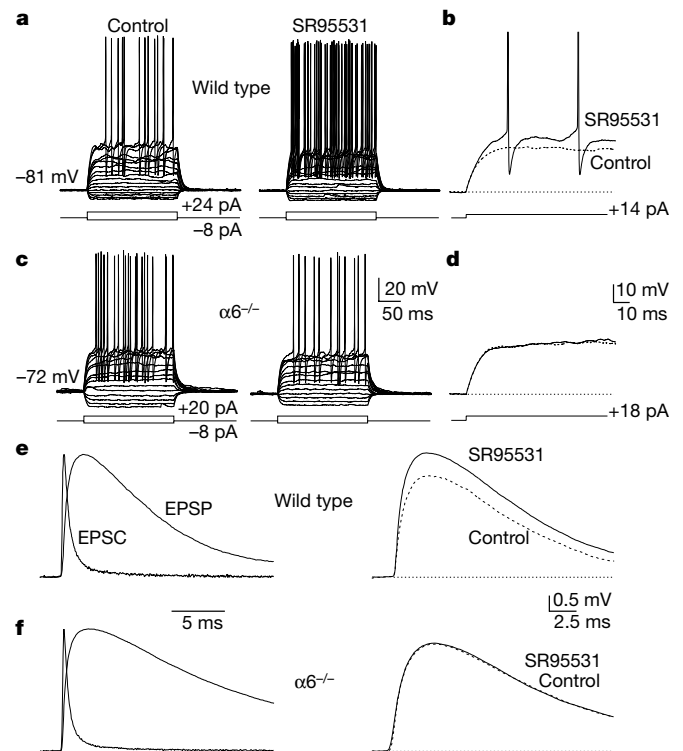


Figure 2 GABA_A receptors and GC excitability. **a**, Voltage recording from a P35 wild-type GC. The current injection protocol (-8 to $+24$ pA in increments of 2 pA) was repeated in the presence of $10 \mu\text{M}$ SR95531. Note the similar resting potential but increased number of action potentials. **b**, Voltage records from a different cell with threshold current injection, in the presence (solid line) and absence (dashed line) of SR95531. Threshold current was reduced from 20 to 14 pA. **c**, **d**, Same as **a** and **b** but from P35 $\alpha 6^{-/-}$ GCs. In these cells, SR95531 did not alter input conductance or excitability. **e**, Left, superimposed normalized average EPSP and EPSC waveforms, recorded from the same GC. Right, traces are average EPSP waveforms taken from a different P35 wild-type GC in the presence and absence of SR95531. **f**, Same as **e** but from P35 $\alpha 6^{-/-}$ GCs. Note the lack of effect of SR95531 on EPSP amplitude (superimposed).

not detect any change in EPSC properties (Table 1), voltage-activated currents (Na^+ current or transient A-type and delayed rectifier K^+ currents; data not shown) or the frequency, pattern and waveform of action potentials in wild-type and $\alpha 6^{-/-}$ GCs (Table 1 and Fig. 2a, c). However, we did detect a difference in the magnitude of a 'standing outward' current. When P35 GCs were held at -20 mV, in the presence of 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX), D(-)-2-amino-5-phosphonopentanoic acid (AP5), SR95531, strychnine and tetrodotoxin (TTX), this current was significantly larger in $\alpha 6^{-/-}$ GCs (43.6 ± 2.0 versus 34.7 ± 2.6 pA/pF; $P < 0.05$, $n = 11$ and 13). A similar standing outward current, identified as a voltage-independent K^+ leak conductance (IK_{SO}), has been reported in cultured cerebellar GCs²³. This conductance has been attributed to the expression of TASK-1 (ref. 11), a member of the growing family of K^+ channels with four transmembrane and two pore-forming domains (2-PK family)¹⁰.

IK_{SO} is a non-inactivating current that shows a quasi-instantaneous decrease in amplitude on hyperpolarization^{11,23}. In our recordings from P35 wild-type (Fig. 3a) and $\alpha 6^{-/-}$ GCs, the characteristics of the standing outward current were similar, and its overall behaviour was consistent with the known properties of 2-PK channels, such as TASK-1 (refs 11, 12). The K^+ permeability of this leak conductance was examined by varying the external K^+ concentration, $[\text{K}^+]_o$, and applying hyperpolarizing voltage ramps. In wild-type cells, increasing $[\text{K}^+]_o$ from 2.5 to 71.3 mM shifted the current reversal by $+65 \pm 5$ mV ($n = 5$), whereas in symmetrical K^+ the current reversed close to 0 mV (0.6 ± 1.6 mV; $n = 3$) and exhibited a linear I - V relationship, characteristic of a voltage-independent K^+ leak conductance. In $\alpha 6^{-/-}$ mice, the shape of the I - V relationship in control conditions was similar to that seen in wild-type P35 GCs, but the magnitude of the current was significantly increased (Fig. 3b). The difference current (Fig. 3c, inset) reversed at about -100 mV, close to the potential predicted for a pure K^+ conductance (arrowed), and showed moderate outward rectification that is well described by the Goldman-Hodgkin-Katz equation. Examination of the relationship between conductance and membrane voltage confirmed that the increased conductance in the $\alpha 6^{-/-}$ GCs was essentially voltage independent, with an average value of 160 pS between -10 and -100 mV ($P < 0.001$,

Kolmogorov-Smirnov test). This behaviour is characteristic of an open rectifier K^+ conductance, such as TASK-1 (refs 9, 11, 12, 24). The increase in leak conductance was specific to GCs, as the current at -20 mV in molecular layer interneurons was not changed (51.0 ± 6.6 versus 49.5 ± 5.3 pA/pF, $n = 9$ wild-type and $n = 11$ $\alpha 6^{-/-}$). Unlike other 2-PK channels that are present in GCs^{10,25,26}, both TASK-1 and TASK-3 are blocked by external acidification^{24,26,27}. Consistent with an involvement of one or both of these channels in the increased leak conductance of $\alpha 6^{-/-}$ GCs, changing the external pH (pH_o) from 7.5 to 6.2 produced a significantly greater reduction in the standing outward current of P35 $\alpha 6^{-/-}$ compared with wild-type GCs ($-35.8 \pm 2.2\%$ $P < 0.001$, versus $-21.6 \pm 5.6\%$, $P < 0.01$ $n = 9$ and 10; difference $P < 0.05$).

As shown in Fig. 4a (top panels), TASK-1 messenger RNA was present at a low level in the GC layer of the cerebellar vermis as early as P7 (arrow), and was found throughout the internal GC layer at P14, with increased expression in the adult. In contrast, TASK-3 expression in the internal GC layer declined after P14 (Fig. 4a, middle panel). During postnatal development, the K^+ leak

Table 1 Properties of P35 wild-type and $\alpha 6^{-/-}$ GCs

	Wild-type	$\alpha 6^{-/-}$
Membrane properties		
Resting potential (mV)	-78.2 ± 0.9	-75.5 ± 1.1
Input conductance (nS)	1.1 ± 0.1	0.8 ± 0.1
Action potentials		
Threshold (mV)	-34.7 ± 1.1	-34.0 ± 1.4
Amplitude (mV)	81.2 ± 1.5	79.9 ± 1.8
10–90% rise time (μs)	223 ± 8	215 ± 7
Half-width (μs)	434 ± 12	425 ± 10
Peak after hyperpolarization (mV)	-67.4 ± 0.8	-67.1 ± 1.5
Action potential threshold current (pA)	25.3 ± 2.7	26.2 ± 6.0
Action potential frequency at +40 pA (Hz)	35.8 ± 9.4	36.1 ± 11.3
Action potential frequency versus current (Hz/pA)	1.89	1.71
Spontaneous EPSCs		
Peak conductance (pS)	227 ± 33	238 ± 25
10–90% Rise time (μs)	180 ± 10	210 ± 10
Decay kinetics (ms)	$\tau_{\text{fast}} 0.6 \pm 0.02$ ($83 \pm 3\%$)	$\tau_{\text{fast}} 0.6 \pm 0.05$ ($87 \pm 4\%$)
	$\tau_{\text{slow}} 6.2 \pm 0.7$	$\tau_{\text{slow}} 4.9 \pm 0.8$
Spontaneous EPSPs		
Amplitude (mV)	2.8 ± 0.3	3.1 ± 0.3
10–90% Rise time (ms)	1.8 ± 0.4	2.5 ± 0.3
Half-width (ms)	13.7 ± 2.2	16.7 ± 3.7

All values are mean $\pm$ s.e.m. Resting membrane potential (zero current potential) was measured within 1 min of establishing whole-cell recording. Input conductance was measured from the steady-state voltage response to a -2 pA current injection. Membrane and action potential properties were determined from 25 wild-type and 26 $\alpha 6^{-/-}$ GCs (except for action potential frequency at +40 pA and action potential frequency versus current, where $n = 5$ and 13, respectively). EPSCs were recorded from five wild-type and four $\alpha 6^{-/-}$ GCs. EPSPs were from seven wild-type and six $\alpha 6^{-/-}$ GCs.

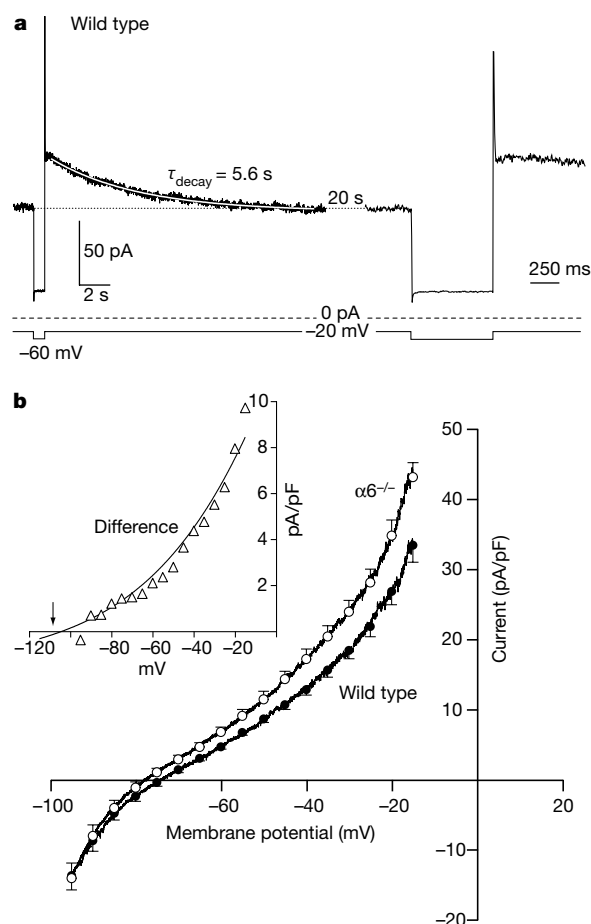


Figure 3 Increased K^+ leak conductance in $\alpha 6^{-/-}$ GCs. **a**, Current from a P35 wild-type GC (with SR95531, AP5, CNQX, strychnine and TTX). Command potential is illustrated below each trace. On stepping to -20 mV, the transient A-type current is followed by a slowly decaying current (single exponential function; white line). After 20 s, a standing outward current persists. The expanded trace shows the quasi-instantaneous inactivation of this current on hyperpolarization. Dotted line shows the amplitude of the persistent current, lower dashed line the zero current. Similar behaviour was seen in $\alpha 6^{-/-}$ GCs. **b**, Averaged I - V relationships from P35 wild-type ($n = 13$) and $\alpha 6^{-/-}$ ($n = 11$) GCs. Symbols and error bars indicate mean and s.e.m. at intervals of 5 mV. The I - V relationships from the two groups were significantly different ($P < 0.05$, Kolmogorov-Smirnov test). Inset, subtracted I - V relationship fitted with the Goldman-Hodgkin-Katz equation. Arrow indicates the predicted reversal potential (-104 mV).

conductance of GCs increased in magnitude. At P14 two populations of GCs were recorded. 'Immature' GCs were characterized by broad, rapidly accommodating action potentials, a low input conductance (0.2 ± 0.02 nS; $n = 16$), and a relatively depolarized resting membrane potential (-53.9 ± 2.1 mV; data not shown). In voltage-clamp recordings from wild-type mice these cells exhibited little standing outward current at -20 mV (6.2 ± 1.1 pA/pF, $n = 8$) and a linear $I-V$ relationship (Fig. 4b). In the same P14 slices,

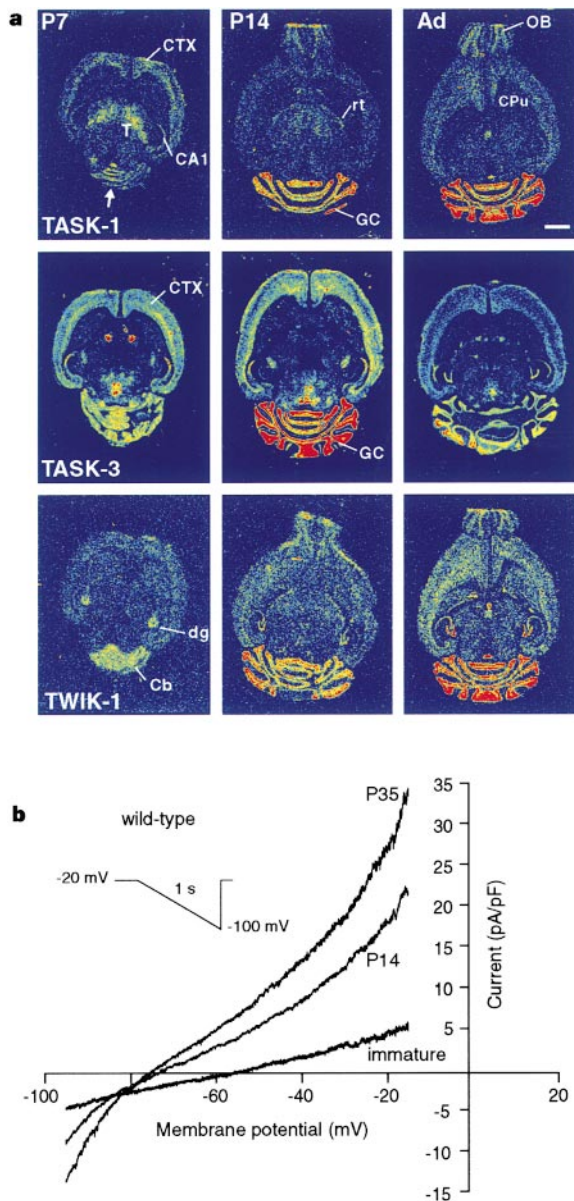


Figure 4 2-PK expression and development of GC leak conductance. **a**, Expression of TASK-1, TASK-3 and TWIK-1 genes during development (false colour images; highest expression in red). TASK-1 mRNA is first evident in developing granule cells of the cerebellar vermis (arrow). By P14, TASK-1 mRNA is found throughout the internal granule cell layer. In adult brain (Ad), the highest TASK-1 expression is in cerebellar GCs. In contrast, TASK-3 is expressed in both internal and external GC layers at P7 but declines after P14. The TWIK-1 gene is pan-neuronal but also shows increased cerebellar expression with age. For each 2-PK gene the pattern of expression in wild-type mice (not shown) was indistinguishable from that in $\alpha 6^{-/-}$ mice. Cb, cerebellum; CPu, caudate-putamen; Ctx, neocortex; dg, dentate granule cells; GC, cerebellar granule cells; OB, olfactory bulb; rt, reticular thalamus; T, thalamus. Scale bar, 1.6 mm. **b**, Averaged $I-V$ relationships (with CNQX, AP5, SR95531, strychnine and TTX) in wild-type GCs ($n = 8-13$) produced by ramp changes in command voltage (inset). Immature GCs were identified by their response to current injection.

'mature' GCs displayed brief non-accommodating action potentials, a higher input conductance (0.6 ± 0.08 nS; $n = 22$), and a resting potential of -71.6 ± 2 mV. These cells exhibited a larger standing outward current (20.5 ± 2.3 pA/pF, $n = 13$) and a doubly rectifying $I-V$ relationship (Fig. 4b). At P35 all GCs showed a more hyperpolarized resting potential and an increased input conductance (Table 1), reflecting the significantly larger K^+ leak conductance (see above). Thus, the expression of TASK-1, but not TASK-3, correlates with the progressive increase in leak conductance observed during GC development. This observation, together with the voltage independence and pH sensitivity of the increased K^+ leak conductance in $\alpha 6^{-/-}$ GCs, suggests that changes in TASK-1 channel activity underlie this increase.

Qualitative comparison of wild-type and $\alpha 6^{-/-}$ cerebellar sections labelled by *in situ* hybridization did not reveal any differences in the expression of the TASK-1 gene. This is perhaps not surprising, as the increased K^+ leak conductance seen in $\alpha 6^{-/-}$ GCs would correspond to the additional contribution of only 20–25 TASK-1 channels per cell (single-channel conductance of 14 pS and a voltage-independent open probability of ~ 0.5 in low $[K^+]_o$; ref. 24). Our data do not allow us to conclusively differentiate between altered channel function or increased gene expression, but the latter is suggested by the results of northern blot analysis of whole cerebella, which showed a 20% increase in TASK-1 mRNA in P35 $\alpha 6^{-/-}$ mice (see Supplementary Information). This change was specific, as mRNA levels for TWIK-1, a weakly inwardly rectifying 2-PK channel¹⁰ present in GCs throughout development (Fig. 4a), did not change. Although additional changes cannot be excluded, the increase in the leak K^+ conductance in $\alpha 6^{-/-}$ GCs was similar to the magnitude of G_{GABA} in wild-type GCs, effectively maintaining a normal input conductance. This equilibration of input conductance would be expected to prevent inappropriate activation of GCs, and may explain the absence of motor deficit in adult $\alpha 6^{-/-}$ mice^{19,21,22}.

Our results show that the GABA_A receptor-mediated tonic conductance in GCs is due exclusively to the activation of receptors containing the $\alpha 6$ subunit. We show that removal of this shunting inhibition in $\alpha 6^{-/-}$ mice leads to an increase in a voltage-independent K^+ leak conductance that maintains normal GC excitability. Although the mechanism of this interaction remains to be determined, the existence of this adaptation highlights the dynamic interplay that occurs between neurotransmitter-gated and other ion channel families^{28,29}. Modulation of 2-PK channels by neurotransmitters, acting through G-protein-coupled receptors, is an important mechanism by which neuronal excitability may be altered in the short term^{11,12}. Our data suggest that these channels, which are widely expressed in the central nervous system^{10,12,24}, can also participate in the long-term homeostatic regulation of neuronal excitability. □

Methods

Parasagittal cerebellar slices (150–250 μ m) were prepared from $\alpha 6^{-/-}$ ($\Delta\alpha 6$ LacZ¹⁹) and wild-type (129/SvJ $\times$ C57BL/6J) mice as described⁸. Slices were perfused at 23–25 °C with a solution containing (in mM): 125 NaCl, 2.5 KCl, 2 CaCl₂, 1 MgCl₂, 26 NaHCO₃, 1.25 NaH₂PO₄, 25 glucose; pH 7.4 when bubbled with 95% O₂ and 5% CO₂. For experiments in which pH_o was changed the external solution contained (in mM): 130 NaCl, 2.5 KCl, 2 CaCl₂, 1 MgCl₂, 10 HEPES; 1.25 NaH₂PO₄, 25 glucose; pH was adjusted using NaOH. Whole-cell recordings were made using Axopatch 200A and B amplifiers (Axon Instruments; Foster City, CA). Cell parameters were measured from amplifier settings: at P7, P14 and P35 whole-cell capacitance (in pF) was 3.1 ± 0.5 ($n = 7$), 2.7 ± 0.1 ($n = 35$) and 2.4 ± 0.2 ($n = 24$) in wild-type GCs, and 3.6 ± 0.5 ($n = 7$), 2.8 ± 0.1 ($n = 48$) and 2.5 ± 0.1 ($n = 30$) in $\alpha 6^{-/-}$ GCs. For voltage-clamp recording of G_{GABA} , the pipette solution contained (in mM) 140 CsCl; 4 NaCl; 0.5 CaCl₂; 10 HEPES; 5 EGTA; 2 Mg-ATP (adjusted to pH 7.3 with CsOH). For all other recordings the pipette solution contained (in mM): 130 K-gluconate; 4 NaCl; 0.5 CaCl₂; 10 HEPES; 5 EGTA; 2 Mg-ATP (adjusted to pH 7.3 with KOH). For this solution, all voltages were corrected for a measured liquid junction potential of -4.8 mV. The following drugs were added to the external solution as indicated: 10 μ M D-AP5, 10 μ M SR95531 (Research Biochemicals), 1 μ M strychnine (Sigma), 5 μ M CNQX (Tocris Cookson), 0.5 μ M TTX (Tocris Cookson).

We carried out data acquisition using pClamp7 or Axograph 4.2 software (Axon). Synaptic events were recorded onto digital audio tape (DTR-1204; BioLogic, Claix, France);

DC to 20 kHz) with the amplifier filter set at 10 kHz. Spontaneous EPSCs were digitized at 33.3 kHz. Spontaneous EPSPs were filtered at 2 kHz and digitized at 10 kHz. Synaptic events and action potentials were analysed either using software written by Stephen Traynelis (Emory University, Atlanta, GA) or Axograph 4.2. Further analysis was performed using Igor Pro 3.14 (WaveMetrics, Lake Oswego, OR). Statistical tests were performed using STATISTICA 5.1 (StatSoft, Tulsa, OK) and considered significant at $P < 0.05$. Differences between groups were examined using the Student's t -test (when measures were normally distributed; Shapiro–Wilk test), the Mann–Whitney U test or Wilcoxon matched-pair test (when distributions were not normal), or the Kolmogorov–Smirnov test, as indicated.

We carried out *in situ* hybridization with ^{35}S -labelled oligodeoxynucleotides using two independent oligonucleotides for each gene, designed to hybridize to different mRNA regions. The specific oligonucleotide sequences were mTWIK-1a, 5'-CTC CAC CAG GCG CAC GCA CGA GCT GCC GGC CAG GGA CTG CAG CAT CTT-3'; mTWIK-1b, 5'-GTG GTC TGC AGA GCC ATC CTC ATA GGG TGG GGA CTG GGA GGC CAC AAA AG-3' (EMBL accession number AF03017); mTASK-1a, 5'-GGT GCA CAC GAT GAG AGC CAA CGT GCG CAC ATT CTG CCG CTT CAT CG-3'; mTASK-1b, 5'-TCA CAC CGA GCT CCT GCG CTT CAT GAG GCC GCG GAA GGC AGC CAG-3' (accession number AB013345); rTASK-3a, 5'-GAT GGA CTT GCG ACG GAT GTG CAG CCT GTG GTT TTC CCC GCA GGT GTG CAT-3'; rTASK-3b, 5'-GTA CTT GCC TCT GAG GCG GAC CTC TTC TGC TTT AAG TTT CTC CTC-3' (accession number AF192366). Each pair of 2-PK probes gave identical results. Images are from a 14-day exposure to X-ray film using the mTWIK-1b, mTASK-1a and rTASK-3a probes.

For northern blot analysis, complementary DNA probes were made by RT–PCR with mouse brain mRNA, and subcloned into pPCR–Script (Stratagene). For mTASK-1, the primers were 5'-AAG GAC CAG GCG CTG CAG AC-3' (554–573; sense) and 5'-TGC ACC GTG CCA AGA GGG-3' (1343–1326; antisense; nucleotide numbers are from accession number AB013345). The 790-bp cDNA fragment was released from pPCR–Script with *Bam*HI and *Sac*I. For mTWIK-1, the primers were 5'-AGC GTG TCA CCG TGC ATG TC-3' (751–770; sense) and 5'-AAT GGA TGC AGT CAA GAC TC-3' (1342–1322, antisense; numbers taken from accession number AF03017). The 591-bp cDNA fragment was released from pPCR–Script with *Bam*HI and *Not*I. Specificity was confirmed on mouse tissues (Clontech) and matched published data. The mTASK-1 probe hybridized to a 4.4-kb band and the mTWIK-1 probe hybridized to a 2.4-kb band. Total RNA was isolated from cerebella of P35 wild-type and $\alpha 6^{-/-}$ mice (eight each). Northern blot analysis was performed using 3 μg of RNA per lane and quantified on a Phosphor imager (Typhoon 8600; Molecular Dynamics). Blots were stripped and normalized with a GAPDH cDNA probe (Ambion).

Received 15 May; accepted 23 October 2000.

- Bernander, O., Douglas, R. J., Martin, K. A. & Koch, C. Synaptic background activity influences spatiotemporal integration in single pyramidal cells. *Proc. Natl Acad. Sci. USA* **88**, 11569–11573 (1991).
- Salin, P. A. & Prince, D. A. Spontaneous GABA_A receptor-mediated inhibitory currents in adult rat somatosensory cortex. *J. Neurophysiol.* **75**, 1573–1588 (1996).
- Häusser, M. & Clark, B. A. Tonic synaptic inhibition modulates neuronal output pattern and spatiotemporal synaptic integration. *Neuron* **19**, 665–678 (1997).
- Rossi, D. J. & Hamann, M. Spillover-mediated transmission at inhibitory synapses promoted by high affinity $\alpha 6$ subunit GABA_A receptors and glomerular geometry. *Neuron* **20**, 783–795 (1998).
- Kaneda, M., Farrant, M. & Cull-Candy, S. G. Whole-cell and single-channel currents activated by GABA and glycine in granule cells of the rat cerebellum. *J. Physiol. (Lond.)* **485**, 419–435 (1995).
- Tia, S., Wang, J. F., Kotchabhakdi, N. & Vicini, S. Developmental changes of inhibitory synaptic currents in cerebellar granule neurons: role of GABA_A receptor $\alpha 6$ subunit. *J. Neurosci.* **16**, 3630–3640 (1996).
- Wall, M. J. & Usowicz, M. M. Development of action potential-dependent and independent spontaneous GABA_A receptor-mediated currents in granule cells of postnatal rat cerebellum. *Eur. J. Neurosci.* **9**, 533–548 (1997).
- Brickley, S. G., Cull-Candy, S. G. & Farrant, M. Development of a tonic form of synaptic inhibition in rat cerebellar granule cells resulting from persistent activation of GABA_A receptors. *J. Physiol. (Lond.)* **497**, 753–759 (1996).
- Duprat, F. *et al.* TASK, a human background K⁺ channel to sense external pH variations near physiological pH. *EMBO J.* **16**, 5464–5471 (1997).
- Lesage, F. *et al.* TWIK-1, a ubiquitous human weakly inward rectifying K⁺ channel with a novel structure. *EMBO J.* **15**, 1004–1011 (1996).
- Millar, J. A. *et al.* A functional role for the two-pore domain potassium channel TASK-1 in cerebellar granule neurons. *Proc. Natl Acad. Sci. USA* **97**, 3614–3618 (2000).
- Talley, E. M., Lei, Q., Sirois, J. E. & Bayliss, D. A. TASK-1, a two-pore domain K⁺ channel, is modulated by multiple neurotransmitters in motoneurons. *Neuron* **25**, 399–410 (2000).
- Eccles, J. C., Llinas, R. & Sasaki, K. The mossy fibre-granule cell relay of the cerebellum and its inhibitory control by Golgi cells. *Exp. Brain Res.* **1**, 82–101 (1966).
- Maex, R. & De Schutter, E. Synchronization of golgi and granule cell firing in a detailed network model of the cerebellar granule cell layer. *J. Neurophysiol.* **80**, 2521–2537 (1998).
- Watanabe, D. *et al.* Ablation of cerebellar Golgi cells disrupts synaptic integration involving GABA inhibition and NMDA receptor activation in motor coordination. *Cell* **95**, 17–27 (1998).
- Saxena, N. C. & Macdonald, R. L. Properties of putative cerebellar γ -aminobutyric acid A receptor isoforms. *Mol. Pharmacol.* **49**, 567–579 (1996).
- Nusser, Z., Sieghart, W. & Somogyi, P. Segregation of different GABA_A receptors to synaptic and extrasynaptic membranes of cerebellar granule cells. *J. Neurosci.* **18**, 1693–703 (1998).
- Sassoe-Pognetto, M., Panzanelli, P., Sieghart, W. & Fritschy, J.-M. Colocalization of multiple GABA_A receptor subtypes with gephyrin at postsynaptic sites. *J. Comp. Neurol.* **420**, 481–498 (2000).
- Jones, A. *et al.* Ligand-gated ion channel subunit partnerships: GABA_A receptor $\alpha 6$ subunit gene inactivation inhibits δ subunit expression. *J. Neurosci.* **17**, 1350–1362 (1997).
- Nusser, Z. *et al.* Alterations in the expression of GABA_A receptor subunits in cerebellar granule cells after the disruption of the $\alpha 6$ subunit gene. *Eur. J. Neurosci.* **11**, 1685–1697 (1999).

- Homanics, G. E. *et al.* Gene knockout of the $\alpha 6$ subunit of the γ -aminobutyric acid type A receptor: lack of effect on responses to ethanol, pentobarbital, and general anesthetics. *Mol. Pharmacol.* **51**, 588–596 (1997).
- Korpi, E. R. *et al.* Cerebellar granule-cell-specific GABA_A receptors attenuate benzodiazepine-induced ataxia: evidence from $\alpha 6$ -subunit-deficient mice. *Eur. J. Neurosci.* **11**, 233–240 (1999).
- Watkins, C. S. & Mathie, A. A non-inactivating K⁺ current sensitive to muscarinic receptor activation in rat cultured cerebellar granule neurons. *J. Physiol. (Lond.)* **491**, 401–412 (1996).
- Leonoudakis, D. *et al.* An open rectifier potassium channel with two pore domains in tandem cloned from rat cerebellum. *J. Neurosci.* **18**, 868–877 (1998).
- Fink, M. *et al.* A neuronal two P domain K⁺ channel stimulated by arachidonic acid and polyunsaturated fatty acids. *EMBO J.* **17**, 3297–3308 (1998).
- Kim, Y., Bang, H. & Kim, D. TASK-3, a new member of the tandem pore K⁺ channel family. *J. Biol. Chem.* **275**, 9340–9347 (2000).
- Lopes, C. M., Gallagher, P. G., Buck, M. E., Butler, M. H. & Goldstein, S. A. Proton block and voltage gating are potassium-dependent in the cardiac leak channel *Kcnk3*. *J. Biol. Chem.* **275**, 16969–16978 (2000).
- Desai, N. S., Rutherford, L. C. & Turrigiano, G. G. Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nature Neurosci.* **2**, 515–520 (1999).
- Turrigiano, G., Abbott, L. F. & Marder, E. Activity-dependent changes in the intrinsic properties of cultured neurons. *Science* **264**, 974–977 (1994).

Acknowledgements

Supported by a Wellcome Trust Programme grant to S.G.C.-C. and a Human Frontier Science Program grant to W.W. We thank L. Cathala, B. Clark, M. Häusser, S.-Q. Liu, T. Takahashi and D. Wyllie for comments on the manuscript, and A. Mathie and B. Robertson for valuable discussion.

Correspondence and requests for materials should be addressed to M.F. (e-mail: m.farrant@ucl.ac.uk).

The lipid phosphatase SHIP2 controls insulin sensitivity

Serge Clément^{*†}, Ulrike Krause^{†‡}, Florence Desmedt[§], Jean-François Tanti^{||}, Jens Behrends^{*}, Xavier Pesesse[§], Takehiko Sasaki[¶], Joseph Penninger[¶], Margaret Doherty[#], Willy Malaisse[#], Jacques E. Dumont^{*§}, Yannick Le Marchand-Brustell^{||}, Christophe Erneux[§], Louis Hue[‡] & Stéphane Schurmans^{*}

^{*}IRIBHN, IBMM, rue des Professeurs Jeener et Brachet 12, 6041 Gosselies, Belgium

[‡]ICP, Hormone and Metabolic Research Unit, Avenue Hippocrate 75, 1200 Brussels, Belgium

[§]IRIBHN, Campus Erasme, route de Lennik 808, 1070 Brussels, Belgium

^{||}INSERM E9911, Faculté de Médecine, Avenue de Valombrose, 06107 Nice, Cedex 02, France

[¶]Agex Institute/Ontario Cancer Institute, Department of Medical Biophysics and Immunology, University of Toronto, University Avenue 620, M5G 2C1 Toronto/Ontario, Canada

[#]Laboratoire de Médecine Expérimentale, Campus Erasme, route de Lennik 808, 1070 Brussels, Belgium

[†]These authors contributed equally to this work

Insulin is the primary hormone involved in glucose homeostasis, and impairment of insulin action and/or secretion has a critical role in the pathogenesis of diabetes mellitus. Type-II SH2-domain-containing inositol 5-phosphatase, or 'SHIP2', is a member of the inositol polyphosphate 5-phosphatase family¹. *In vitro* studies have shown that SHIP2, in response to stimulation by numerous growth factors and insulin, is closely linked to signalling events mediated by both phosphoinositide-3-OH kinase and Ras/mitogen-activated protein kinase^{2–5}. Here we report the generation of mice lacking the SHIP2 gene. Loss of SHIP2 leads to increased sensitivity to insulin, which is characterized by severe neonatal hypoglycaemia, deregulated expression of the genes involved in gluconeogenesis, and perinatal death. Adult mice that are heterozygous for the SHIP2 mutation have increased glucose tolerance and insulin sensitivity associated with an increased recruitment of the GLUT4 glucose transporter and

increased glycogen synthesis in skeletal muscles. Our results show that SHIP2 is a potent negative regulator of insulin signalling and insulin sensitivity *in vivo*.

In contrast to the related 5-phosphatase SHIP, whose expression is limited to the haematopoietic system, SHIP2 is expressed in a wide range of tissues¹⁻⁷. To characterize SHIP2 functions *in vivo*, we generated and analysed a mouse strain deficient in the SHIP2 gene (Fig. 1). After electroporation of embryonic stem (ES) cells with the targeting vector, we used Southern blotting with two different probes and restriction enzymes to identify recombinant clones with a 7.3-kilobase (kb) genomic DNA deletion on one allele of the SHIP2 gene (Fig. 1a, b). Crossing of chimaeric males with C57BL/6 females resulted in F₁ heterozygous (SHIP2^{+/-}) mice that have no obvious abnormalities. The body weight at 8 weeks, life expectancy, and spontaneous or irradiation-induced tumour incidence were not significantly different from wild-type SHIP2^{+/+} mice. F₁ SHIP2^{+/-} mice were then intercrossed and 182 viable offspring were genotyped at birth; of these, 47 (26%) were SHIP2^{+/+}, 94 (51%) SHIP2^{+/-} and 41 (22%) SHIP2^{-/-}. These frequencies were within mendelian expectations for transmission of an autosomal gene, and suggest that disruption of both SHIP2 alleles does not cause embryonic lethality. Mouse embryonic fibroblasts (MEFs) derived from day-13.5 embryos were analysed to confirm the reduced levels and the complete absence of SHIP2 messenger RNA (Fig. 1c) or protein (Fig. 1d) in SHIP2^{+/-} and SHIP2^{-/-} mice, respectively.

At birth, SHIP2^{-/-} mice were phenotypically indistinguishable from their littermates; most of them were able to feed, although progressively less efficiently than SHIP2^{+/+} or SHIP2^{+/-} mice. Close monitoring revealed that SHIP2^{-/-} mice became progressively cyanotic or pale and lethargic in the first 24 h of life. Some newborn

SHIP2^{-/-} pups presented signs of respiratory distress and all failed to gain weight (SHIP2^{+/+} 1.60 ± 0.02 g, SHIP2^{+/-} 1.63 ± 0.04 g, SHIP2^{-/-} 1.20 ± 0.01 g, for a typical litter of 10 newborns, 18 h after birth; Student's *t*-test, *P* < 0.005), and died within 3 days after birth. The cause of the respiratory distress observed in some of the SHIP2^{-/-} mice was not a lack of surfactant or an abnormal differentiation of surfactant-producing cells, as the expression of surfactant-associated proteins (SP-A, SP-B and SP-C) and of TTF-1 or C/EBP-α transcription factors was normal in the lungs of SHIP2^{-/-} mice (data not shown). Moreover, haematoxylin and eosin staining of SHIP2^{-/-} lung, as well as of brain, heart, thymus, liver, stomach, pancreas, kidney, skin, muscle, spleen, bladder, and small and large intestine sections revealed no particular abnormalities (see Supplementary Information).

In newborns, severe hypoglycaemia is often associated with cyanotic episodes, apnea, respiratory distress, refusal to feed and somnolence⁸. No difference in blood glucose concentrations among the different genotypes was observed 2 and 8 h after delivery (data not shown). When analysed between 10 to 15 h postpartum, however, blood glucose concentration in SHIP2^{-/-} mice was significantly lower than in SHIP2^{+/-} and SHIP2^{+/+} mice (Fig. 2a; Student's *t*-test, *P* < 10⁻⁶). Hypoglycaemia was not caused by glycosuria (data not shown), nor by excessive secretion of insulin by the pancreatic β-cells: plasma insulin levels were significantly lower in SHIP2^{-/-} mice than in SHIP2^{+/-} and SHIP2^{+/+} mice (Fig. 2a; Student's *t*-test, *P* < 0.05). Histology and immunohistochemistry of pancreatic islets using anti-insulin, anti-glucagon and anti-somatostatin antibodies did not reveal any abnormalities of antigen expression and distribution in SHIP2^{-/-} mice (data not shown). To test whether the hypoglycaemia was the cause of postnatal death, we injected SHIP2^{-/-} mice with D-glucose. Hypoglycaemic SHIP2^{-/-}

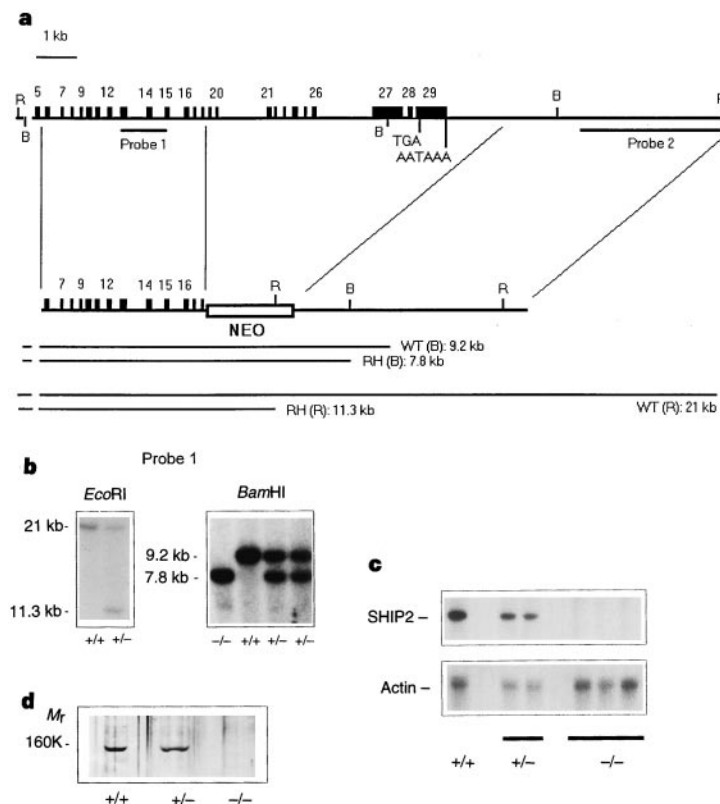


Figure 1 Targeted disruption of *SHIP2* gene. **a**, The wild-type allele (top) and the targeting vector (bottom) are represented (exons, filled boxes), as well as the two probes used in Southern analysis and the DNA fragments generated after digestion with *EcoRI* (R) and *BamHI* (B). **b**, Southern blot analysis of wild-type (+/+) and recombinant (+/-) ES cells, and of progeny of a heterozygote cross. **c**, Northern blot analysis of mRNA (2 µg per lane)

isolated from SHIP2^{+/+}, SHIP2^{+/-} and SHIP2^{-/-} MEFs hybridized with a mouse SHIP2 or actin cDNA fragment as the probe. **d**, Western blot analysis of proteins (100 µg per lane) isolated from SHIP2^{+/+}, SHIP2^{+/-} and SHIP2^{-/-} MEFs probed with a rabbit anti-SHIP2 antibody.

mice were transiently rescued for up to 96 h by repeated injections of D-glucose during the first 24 h postpartum (Fig. 2b). About 10 min after each injection, SHIP2^{-/-} newborns recovered a normal skin colour and were less lethargic than non-injected mice. Prolonged survival was also observed when SHIP2^{-/-} mice were injected with a neutralizing antibody to insulin within the first hour after birth (Fig. 2b). In addition, anti-insulin antibody injection of SHIP2^{-/-} mice increased glucose concentrations significantly at 10–15 h postpartum as compared with non-injected mice (Fig. 2a; Student's *t*-test, *P* < 0.0002). These data indicate that hypoglycaemia in SHIP2^{-/-} mice is not caused by an increased production of insulin, but rather results from an increased sensitivity to insulin.

The liver has a principal role in glucose homeostasis at birth: it provides glucose to the blood through gluconeogenesis^{9,10}. During that critical period, transcription of several hepatic enzymes involved in gluconeogenesis is activated in response to hormonal and dietary conditions^{9–11}. Interfering with glucagon or insulin signalling cascades at or just before birth results in delayed or premature appearance of these enzymes, and in neonatal hypoglycaemia or diabetes^{9,12–15}. Expression of phosphoenolpyruvate carboxy-kinase (PEPCK), a key enzyme of gluconeogenesis, was very low or absent in liver of SHIP2^{-/-} mice, as compared to SHIP2^{+/+} mice (Fig. 2c). Injection of SHIP2^{-/-} mice with a neutralizing anti-

insulin antibody restored a normal expression of PEPCK mRNA (Fig. 2c). Expression of hepatic tyrosine aminotransferase (TAT-5') and glucose-6-phosphatase (G-6-Pase), two other gluconeogenic enzymes induced after birth, were also decreased, albeit to a lesser extent than PEPCK (Fig. 2c). Levels of C/EBP- α , C/EBP- β and aldolase B mRNA were unaffected by the mutation (data not shown). Thus, the absence of SHIP2 leads to decreased expression of several key gluconeogenic enzymes, contributing to hypoglycaemia in newborn SHIP2^{-/-} mice. Together, these data indicate that SHIP2^{-/-} liver cells have enhanced sensitivity to insulin *in vivo*.

As SHIP2 is a critical negative regulator of insulin sensitivity *in vivo*, and as loss of SHIP2 leads to lethal hypoglycaemia in newborn mice, we investigated whether decreased amounts of SHIP2 expression in SHIP2^{+/-} mice (Figs 1c, d, 3c) would alter insulin sensitivity. There was no significant difference in basal blood glucose or plasma insulin levels between adult SHIP2^{+/+} and SHIP2^{+/-} mice, either after an overnight fasting (Fig. 3a), or when freely fed (glucose levels in freely fed mice, 9.8 ± 0.5 mM versus 8.8 ± 0.2 mM in SHIP2^{+/+} and SHIP2^{+/-} mice, respectively; insulin levels in freely fed mice, 1.17 ± 0.29 μ g l⁻¹ versus 0.96 ± 0.16 μ g l⁻¹ in SHIP2^{+/+} and SHIP2^{+/-} mice, respectively). However, injection of D-glucose resulted in a more rapid glucose clearance in SHIP2^{+/-} than in SHIP2^{+/+} mice: glycaemia was significantly lower at all time points

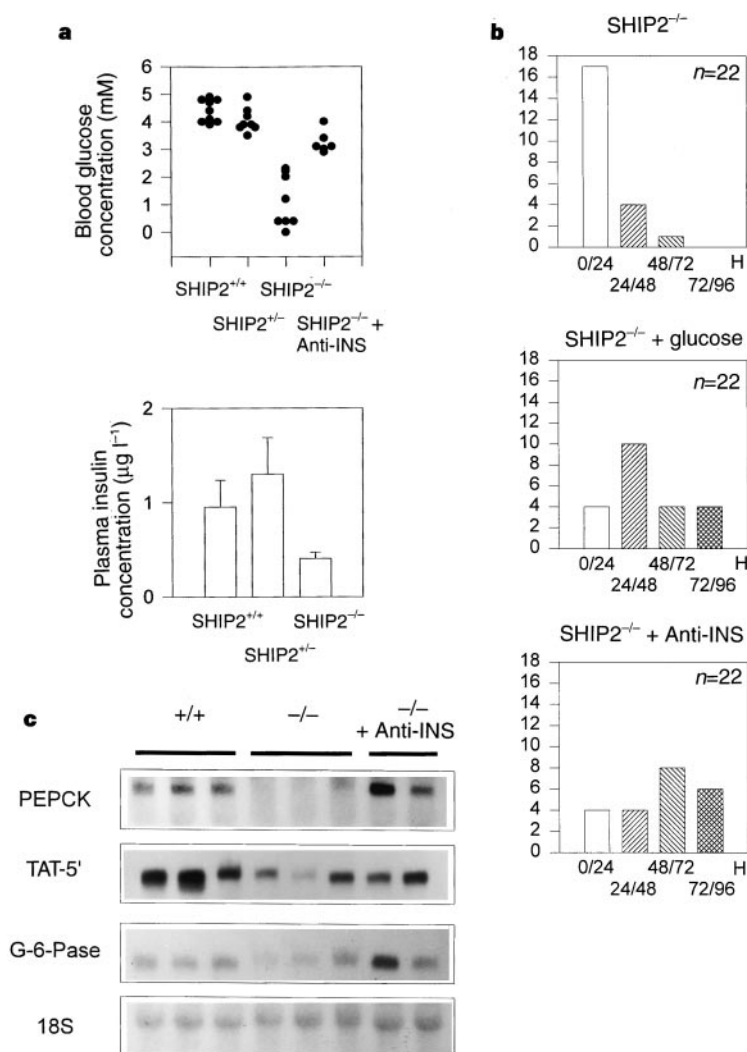


Figure 2 Impaired glucose homeostasis in SHIP2^{-/-} newborns. **a**, Blood glucose (top) and plasma insulin (bottom) concentrations in SHIP2^{+/+}, SHIP2^{+/-} and SHIP2^{-/-} newborn mice. Values are expressed as the means $\pm$ s.e.m. from the analysis of nine SHIP2^{+/+}, eight SHIP2^{+/-}, eight SHIP2^{-/-} and six anti-insulin antibody-treated SHIP2^{-/-} mice. **b**, Mortality of SHIP2^{-/-} newborns injected or not with either D-glucose or a neutralizing

guinea-pig anti-insulin antibody. Mann–Whitney tests indicated that mortality of non-treated and treated SHIP2^{-/-} mice was significantly different (*P* < 0.001). **c**, Northern blot analysis of PEPCK, TAT-5' and G-6-Pase expression in liver of between 2- and 3.5-hour-old SHIP2^{+/+} and SHIP2^{-/-} newborn mice.

in SHIP2^{+/-} mice (Fig. 3a). The increased glucose clearance in SHIP2^{+/-} mice was not a consequence of an increased release of insulin: 30 min after glucose administration, insulin levels were also significantly lower in SHIP2^{+/-} than in wild-type mice (Fig. 3a). Moreover, the output of insulin from isolated pancreatic islets in response to glucose was not significantly different in SHIP2^{+/-} and SHIP2^{+/-} mice (see Supplementary Information). Insulin hyper-

sensitivity was shown when mice were injected with insulin, that is, a significantly more profound hypoglycaemia was observed by 30 and 60 min after injection in SHIP2^{+/-} mice than in SHIP2^{+/+} mice (Fig. 3b).

Insulin stimulates glucose transport and storage of glucose as glycogen into skeletal muscles, through translocation of the GLUT4 glucose transporter from intracellular stores to the cell surface¹⁶, and

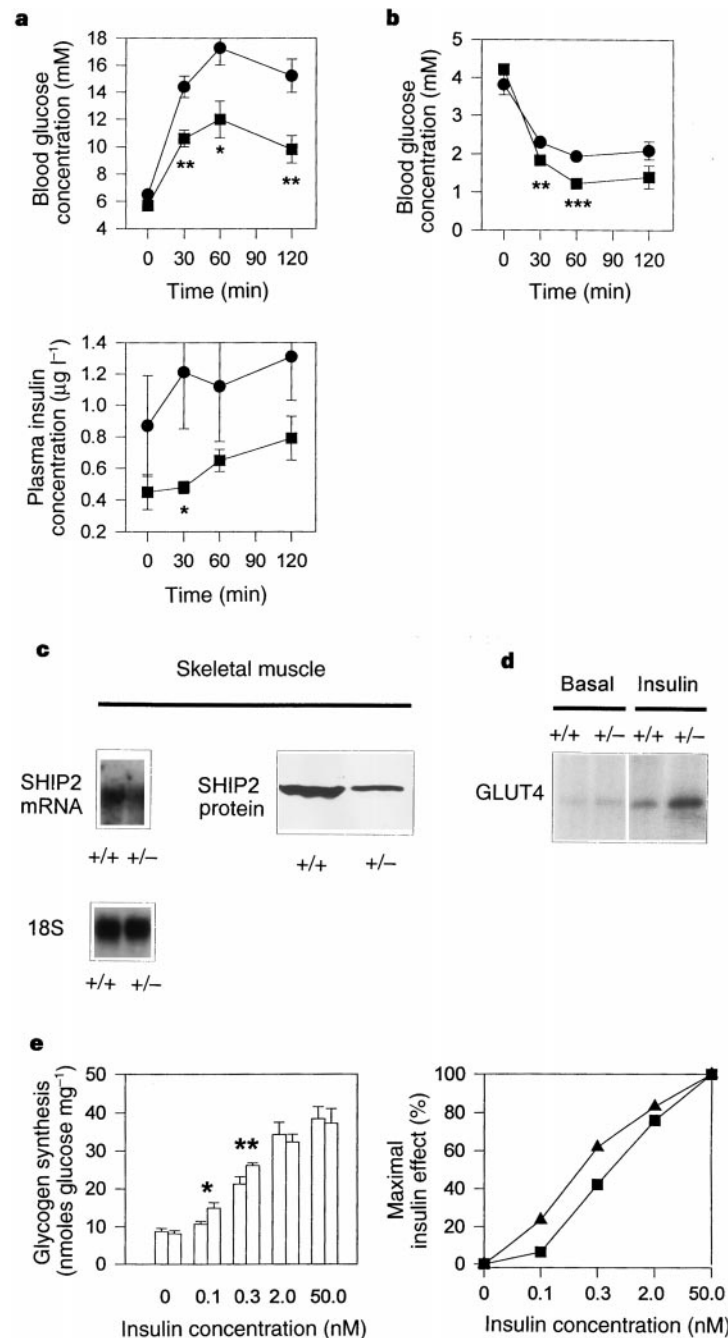


Figure 3 Increased insulin sensitivity in adult SHIP2^{+/-} mice. **a**, Blood glucose (top) and plasma insulin (bottom) concentrations in a glucose tolerance test. Values are expressed as the means $\pm$ s.e.m. from six SHIP2^{+/+} (filled circles) and SHIP2^{+/-} (filled squares) mice that have indistinguishable body weight. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$ (Student's *t*-test). **b**, Blood glucose concentrations in an insulin tolerance test. Values are expressed as the mean $\pm$ s.e.m. obtained from ten SHIP2^{+/+} (filled circles) and SHIP2^{+/-} (filled squares) mice. Two asterisks, $P < 0.01$; three asterisks, $P < 0.001$ (Student's *t*-test). **c**, Northern and Western blot analysis of total RNA and proteins isolated from skeletal muscles of 6–8-week-old SHIP2^{+/+} and SHIP2^{+/-} mice. **d**, GLUT4 expression in skeletal muscles from adult SHIP2^{+/+} and SHIP2^{+/-} mice. Results are representative of

three separate experiments. **e**, Glycogen synthesis in isolated soleus muscles from adult SHIP2^{+/+} and SHIP2^{+/-} mice. Left, values are expressed as nmol glucose incorporated into glycogen per mg of muscle protein and are presented as means $\pm$ s.e.m. of 7–8 muscles (except at 0.1 nM insulin where four muscles were used). At each insulin concentration, the left column represents SHIP2^{+/+} mice, and the right column SHIP2^{+/-} mice. Asterisk, $P < 0.02$; two asterisks, $P < 0.03$ (Student's *t*-test for unpaired data) when comparing SHIP2^{+/+} and SHIP2^{+/-} mice. Right, results are expressed as percentage of the maximum insulin effect (determined by dividing the increment owing to insulin at each hormone concentration by the maximal increment in insulin effect at 50 nM). Filled squares, SHIP2^{+/+}; filled triangles, SHIP2^{+/-}.

also activates glycogen synthase. Loss of one SHIP2 allele resulted in reduced SHIP2 mRNA and protein expression in skeletal muscle cells (Fig. 3c). After an overnight fasting, the amount of GLUT4 transporter in the myocyte plasma membrane fraction was low but similar in SHIP2^{+/+} and SHIP2^{+/-} mice (Fig. 3d). When SHIP2^{+/+} and SHIP2^{+/-} mice were loaded with insulin, however, plasma membrane GLUT4 levels were higher in SHIP2^{+/-} skeletal muscles. Glycogen synthesis in response to insulin stimulation, which reflects both glucose uptake and glycogen synthase activation, was analysed in isolated soleus muscles from SHIP2^{+/+} and SHIP2^{+/-} mice (Fig. 3e). Under basal conditions or in the presence of saturating insulin concentrations (2 and 50 nM), a similar glycogen synthesis was observed in SHIP2^{+/+} and SHIP2^{+/-} muscles. However, stimulation with low, physiological insulin concentrations (0.1 or 0.3 nM) resulted in significantly higher glycogen synthesis in SHIP2^{+/-} muscles than in SHIP2^{+/+} muscles (Fig. 3e, left panel). When expressed as a percentage of the maximum insulin effect, the insulin dose–response curve was shifted towards the lower insulin concentrations in SHIP2^{+/-} mice, as compared to SHIP2^{+/+} mice, reaching a 3.6-fold and a 1.5-fold increase in glycogen synthesis in response to 0.1 and 0.3 nM insulin, respectively (Fig. 3e, right panel). Together, our data indicate that insulin sensitivity is significantly increased in skeletal muscles from heterozygous mice expressing a reduced amount of SHIP2 protein.

From histological analysis performed on SHIP2^{-/-} and SHIP2^{+/+} organs, from tumour incidence in adult SHIP2^{+/-} and SHIP2^{+/+} mice, and from SHIP2^{-/-} and SHIP2^{+/+} haematopoietic cell analysis, we were unable to confer an important role for SHIP2 *in vivo* in the control of systems other than insulin signalling (see Supplementary Information). This restricted role of SHIP2 observed in SHIP2 mutant mice may result from an intracellular compartment-specific action of the enzyme^{17,18}, and/or from the activation of related enzymes that are more specifically associated with growth factors signalling cascade.

The incidence of adult-onset diabetes mellitus has increased markedly and the disease is a major health-care problem¹⁹. Resistance to the stimulatory effect of insulin on glucose use is a key pathogenic feature of most forms of adult-onset (type II, or non-insulin-dependent) diabetes, and contributes to the morbidity associated with autoimmune (type I, or insulin-dependent) diabetes²⁰. It is crucial to have a better understanding of the molecular mechanisms that regulate insulin signalling. Our data in genetically modified mice identify SHIP2 as a critical and essential negative regulator of insulin signalling and insulin sensitivity *in vivo*. Thus, SHIP2 is a potential therapeutic target for the treatment of type II diabetes, and a candidate gene that predisposes to the same disease. □

Methods

Production of SHIP2-deficient mice and genotyping

The targeting vector was constructed by replacing a 7.3-kb genomic fragment containing exons 19–29 and the polyadenylation signal of the mouse SHIP2 gene by a neomycin resistance cassette⁶. The cassette was flanked by 4.0-kb and 5.5-kb mouse genomic DNA fragments at the 5' and 3' regions, respectively. R1 ES cells were electroporated with the targeting plasmid linearized by *SalI*. Homologous recombination at the SHIP2 locus was confirmed by Southern blotting, and SHIP2^{-/-} ES cells were aggregated with morulae derived from CD1 mice. The resulting chimaeric mice transmitted the mutant allele to the progeny.

Northern and western analysis

Messenger RNA was extracted from MEFs using a FastTrack kit (Invitrogen), and total RNA was purified from newborn liver and from adult skeletal muscles using the RNeasy Mini Kit (Qiagen). Two micrograms per lane of mRNA or 20 µg per lane of total RNA were loaded on a 1% agarose gel and transferred to a nylon membrane. The membranes were hybridized with mouse complementary DNA fragments coding for SHIP2, TAT-5', C/EBPα, C/EBPβ, aldolase B, actin or G-6-Pase as probes. Antisense oligonucleotide probes were used for PEPCK mRNA and 18S RNA detection²¹. Supernatants of muscle or MEF homogenates isolated, respectively, from adult SHIP2^{+/-} and SHIP2^{+/+} mice, or from SHIP2^{+/+}, SHIP2^{+/-} and SHIP2^{-/-} day-13.5 embryos were analysed by western blot. Proteins (100 µg per lane) were separated by SDS–PAGE and transferred to nitrocellulose sheets. Saturation, incubation with rabbit anti-SHIP2 antiserum and ECL detection were

done as described²². For *in vivo* GLUT4 expression, SHIP2^{+/+} and SHIP2^{+/-} mice (6–10-week-old) were overnight fasted, anaesthetized and intravenously injected or not with 1 mU per g (body weight) insulin. After 5 min, the skeletal muscles from the hind limbs were removed. A plasma membrane-rich fraction and a total lysate were prepared from myocytes as described^{23,24}. Aliquots of proteins (100 µg) were separated by SDS–PAGE and blotted with anti-GLUT4 antibody. Immune complexes were detected by using ¹²⁵I-labelled protein-A.

Glucose and insulin tolerance tests

SHIP2^{+/+} and SHIP2^{+/-} mice (6–10-week-old) were fasted for more than 12 h and intraperitoneally injected with either 1.5 mg per g (body weight) D-glucose or 1 mU per g (body weight) insulin (Actrapid, Novo Nordisk, Denmark). Blood samples were drawn from the retro-orbital sinus at different time points. Blood glucose concentration was determined enzymatically. Plasma insulin levels were determined using a rat insulin ELISA kit (Merckodia AB, Uppsala, Sweden). Methods for the output of insulin by islet β-cells, and the specific action of SHIP2 on the insulin signalling cascade are given in Supplementary Information.

Treatment with glucose or anti-insulin antibodies

Newborns received either repeated intraperitoneal injections of D-glucose (7%, 100 µl per injection) during the first 24 h after birth, or a single intraperitoneal injection of a guinea-pig anti-porcine insulin polyclonal antibody (1/800 in 0.9% NaCl, 100 µl; ref AB1295, Chemicon Int. Inc., Temecula, CA) in the first postnatal hour. The neutralizing effect of this anti-insulin antibody was shown after injection in adult mice loaded with glucose: significantly increased blood glucose levels were observed after 30 and 60 min, as compared to non-treated or normal guinea-pig serum-treated mice (data not shown).

Analysis of glycogen synthesis in soleus muscles

The two soleus muscles were isolated from overnight fasted SHIP2^{+/+} or SHIP2^{+/-} mice, and processed as described²⁵.

Received 5 July; accepted 23 October 2000.

- Pesesse, X., Deleu, S., De Smedt, F., Drayer, L. & Erneux, C. Identification of a second SH2-domain-containing protein closely related to the phosphatidylinositol polyphosphate 5-phosphatase SHIP. *Biochem. Biophys. Res. Commun.* **239**, 697–700 (1997).
- Habib, T., Hejima, J. A., Moses, R. E. & Decker, S. J. Growth factors and insulin stimulate tyrosine phosphorylation of the 51C/SHIP2 protein. *J. Biol. Chem.* **273**, 18605–18609 (1998).
- Pesesse, X. *et al.* The SH2 domain containing inositol 5-phosphatase SHIP2 displays phosphatidylinositol 3,4,5-triphosphate and inositol 1,3,4,5-tetrakisphosphate 5-phosphatase activity. *FEBS Lett.* **437**, 301–303 (1998).
- Wisniewski, D. *et al.* A novel SH2-containing phosphatidylinositol 3,4,5-triphosphate 5-phosphatase (SHIP2) is constitutively tyrosine phosphorylated and associated with src homologous and collagen gene (SHC) in chronic myelogenous leukemia progenitor cells. *Blood* **93**, 2707–2720 (1999).
- Ishihara, H. *et al.* Molecular cloning of rat SH2-containing inositol phosphatase 2 (SHIP2) and its role in the regulation of insulin signaling. *Biochem. Biophys. Res. Commun.* **260**, 265–272 (1999).
- Schurmans, S. *et al.* The mouse SHIP2 (Inp11) gene: complementary DNA, genomic structure, promoter analysis, and gene expression in the embryo and adult mouse. *Genomics* **62**, 260–271 (1999).
- Liu, Q. *et al.* The SH2-containing inositol polyphosphate 5-phosphatase, SHIP, is expressed during hematopoiesis and spermatogenesis. *Blood* **91**, 2753–2759 (1998).
- Lteif, A. N. & Schwenk, W. F. Hypoglycemia in infants and children. *Endocrinol. Metab. Clin. North Am.* **28**, 619–646 (1999).
- Girard, J., Ferré, P., Pégrier, J.-P. & Dué, P.-H. Adaptations of glucose and fatty acid metabolism during perinatal period and suckling-weaning transition. *Physiol. Rev.* **72**, 507–562 (1992).
- Girard, J. & Pégrier, J.-P. An overview of early post-partum nutrition and metabolism. *Biochem. Soc. Trans.* **26**, 69–74 (1998).
- Jones, C. T. (ed) *The Development of the Metabolism in the Fetal Liver* (Elsevier Biomedical Press, Elsevier, 1982).
- Girard, J., Caquet, D., Bal, D. & Guillet, I. Control of rat liver phosphorylase, glucose-6-phosphatase and phosphoenolpyruvate carboxykinase activities by insulin and glucagon during the perinatal period. *Enzyme* **15**, 272–285 (1973).
- Yeung, D. & Oliver, I. T. Factors affecting the premature induction of phosphoenolpyruvate carboxykinase in neonatal rat liver. *Biochem. J.* **108**, 325–331 (1968).
- Hanson, R. W., Fisher, L., Ballard, F. J. & Reshef, L. The regulation of phosphoenolpyruvate carboxykinase in fetal rat liver. *Enzyme* **15**, 97–100 (1973).
- Benvenisty, N. *et al.* in *Lessons from Animal Diabetes* (eds Shafir, E. A. & Renold, A. E.) 717–733 (Libbey, London, 1984).
- Czech, M. & Corvera, S. Signaling mechanisms that regulate glucose transport. *J. Biol. Chem.* **274**, 1865–1868 (1999).
- Nave, B. T. *et al.* Compartment-specific regulation of phosphoinositide 3-kinase by platelet-derived growth factor and insulin in 3T3-L1 adipocytes. *Biochem. J.* **318**, 55–60 (1996).
- Corvera, S. & Czech, M. P. Direct targets of phosphoinositide 3-kinase products in membrane traffic and signal transduction. *Trends Cell Biol.* **8**, 442–446 (1998).
- Warram, J. H., Rich, S. S. & Krolewski, A. S. in *Joslin's Diabetes Mellitus* (eds Kahn, C. R. & Weir, G. C.) 201–216 (Lea and Febiger, Philadelphia, 1995).
- Kahn, B. Type 2 diabetes: when insulin secretion fails to compensate for insulin resistance. *Cell* **92**, 593–596 (1998).
- Lee, Y.-H., Sauer, B., Johnson, P. F. & Gonzalez, F. J. Disruption of the C/EBPα gene in adult mouse liver. *Mol. Cell. Biol.* **17**, 6014–6022 (1997).
- Bruyns, C., Pesesse, X., Moreau, C., Blero, D. & Erneux, C. The two SH2-domain-containing inositol 5-phosphatases SHIP1 and SHIP2 are coexpressed in human T lymphocytes. *Biol. Chem.* **380**, 969–974 (1999).
- Simpson, I. A. *et al.* Insulin-stimulated translocation of glucose transporters in the isolated rat adipose cells. *Biochim. Biophys. Acta* **763**, 393–407 (1983).

24. Higaki, Y. *et al.* Insulin receptor substrate-2 is not necessary for insulin- and exercise-stimulated glucose transport in skeletal muscle. *J. Biol. Chem.* **274**, 20791–20795 (1999).
25. Stenbit, A. E. *et al.* Diverse effects of Glut 4 ablation on glucose uptake and glycogen synthesis in red and white skeletal muscle. *J. Clin. Invest.* **98**, 629–634 (1996).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank B. Payraastre, G. Giuriato, and J. and R. Merino for sharing unpublished results; E. Marion, T. Grémeaux and L. Maisin for technical assistance; A. Nagy for the R1 ES cells; P. Johnson for rat C/EBP β cDNA clone; and G. Schütz for mouse TAT-5', G-6-Pase, C/EBP α and aldolase B cDNA clones. This work and U.K. were supported by the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Federal Service for Science, Technology and Culture, the Fonds de la Recherche Scientifique Médicale de Belgique, Biomed 2 program, Association contre le Cancer, and Télévie. S.C. is a fellow of the FRIA; F.D. and S.S. are Chargé de Recherche and Chercheur Qualifié of the Belgian FNRS, respectively. J.B. was supported by the Deutsche Forschungsgemeinschaft. Y.L.M.B. and J.F.T. are supported by Institut National de la Santé et de la Recherche Médicale (France).

Correspondence and requests for materials should be addressed to S.S. (e-mail: sschurma@ulb.ac.be).

ICOS co-stimulatory receptor is essential for T-cell activation and function

Chen Dong^{*†}¶, Amy E. Juedes[‡]§, Ulla-Angela Temann^{*§}¶, Sujan Shrestha^{||}, James P. Allison^{||}, Nancy H. Ruddle[‡] & Richard A. Flavell^{*¶}

^{*} Howard Hughes Medical Institute, Section of Immunobiology, and

[¶] Yale University School of Medicine, New Haven, Connecticut 06520, USA

[‡] Department of Epidemiology and Public Health and Section of Immunobiology, Yale University School of Medicine, New Haven, Connecticut 06520, USA

^{||} Howard Hughes Medical Institute, Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA

[§] These authors contributed equally to this work

T-lymphocyte activation and immune function are regulated by co-stimulatory molecules. CD28, a receptor for B7 gene products, has a chief role in initiating T-cell immune responses^{1,2}. CTLA4, which binds B7 with a higher affinity, is induced after T-cell activation and is involved in downregulating T-cell responses^{3,4}. The inducible co-stimulatory molecule (ICOS), a third member of the CD28/CTLA4 family, is expressed on activated T cells^{5,6}. Its ligand B7H/CD28/CTLA4 family, is expressed on B cells and in non-immune tissues after infection of lipopolysaccharide into animals^{6,7}. To understand the role of ICOS in T-cell activation and function, we generated and analysed ICOS-deficient mice. Here we show that T-cell activation and proliferation are defective in the absence of ICOS. In addition, ICOS^{-/-} T cells fail to produce interleukin-4 when differentiated *in vitro* or when primed *in vivo*. ICOS is required for humoral immune responses after immunization with several antigens. ICOS^{-/-} mice showed greatly enhanced susceptibility to experimental autoimmune encephalomyelitis, indicating that ICOS has a protective role in inflammatory autoimmune diseases.

To investigate the role(s) of ICOS in T-cell activation and during *in vivo* immune reactions, we created and analysed ICOS-deficient mice. A gene-targeting construct was generated to replace the three exons that encode murine ICOS protein (except the leader peptide) with a neomycin-resistance gene driven by a PGK promoter

(Fig. 1a). This construct was transfected into embryonic stem (ES) cells, and two ES clones identified to have undergone homologous recombination were used to create ICOS^{+/-} mice by standard approaches. Interbreedings of heterozygous ICOS knockout mice yielded homozygous mice as identified by Southern blot analysis (data not shown). Lymph-node cells from ICOS^{+/-}, ICOS^{-/-} and ICOS^{-/-} mice were activated by anti-CD3 plus CD28 antibodies, and ICOS expression was analysed by polymerase chain reaction with reverse transcription RT-PCR (Fig. 1b) or staining with an ICOS antibody that we developed (Fig. 1c). ICOS^{-/-} cells completely lack ICOS expression (Fig. 1b, c), but retain CTLA4 (Fig. 1b) and CD28 expression (data not shown).

T lymphocytes developed normally in the thymus in the absence of ICOS (data not shown). Furthermore, ICOS knockout mice exhibited normal CD4 and CD8 cell populations in lymph nodes and spleen (data not shown). This indicates that ICOS is not required for T-cell development. To test whether ICOS-deficient cells function normally, we activated lymph-node and splenic cells from ICOS^{+/-}, ICOS^{+/-} and ICOS^{-/-} littermates with different doses of plate-bound anti-CD3. In this culture, co-stimulation of T cells by B7 family members, including B7H/B7RP-1, is provided by B cells and other antigen-presenting cells (APCs). ICOS^{-/-} cells were defective in proliferation after activation when measured by [³H]thymidine incorporation (Fig. 2a). When incubated with wild-type APCs, purified naive CD4⁺ T cells from the knockout mice also exhibited reduced proliferation after anti-CD3 activation (data not shown). Consistent with the reduction in proliferation, there was a significantly higher number of CD4⁺ ICOS^{-/-} cells that retained the naive phenotype, in other words, that were CD62L^{high} after 2 days of activation (Fig. 2b). The hallmark of T-cell activation is the expression of interleukin (IL)-2, which functions as a growth factor to mediate T-cell clonal expansion. ICOS^{-/-} CD4⁺ T cells produced greatly reduced amounts of this cytokine after activation (Fig. 2c). Addition of exogenous IL-2 could rescue most of the proliferative defect of the knockout cells (Fig. 2d). These data indicate that, at least *in vitro*, ICOS has a critical and non-dispensable role in T-cell activation and proliferation.

Previous studies that used anti-ICOS antibodies, soluble B7RP-1 or ICOS-immunoglobulin molecules suggested that ICOS may regulate T-helper cell differentiation and cytokine production^{5,6,9}. To assess whether ICOS is essential for T-helper cell differentiation or effector function, we purified naive CD4⁺ T cells from wild-type and knockout mice and activated them by anti-CD3 and irradiated wild-type APCs. Exogenous recombinant IL-2 was added to compensate for the proliferative defect of the knockout cells. Four days later, the differentiated cells were re-stimulated, and T-helper cell type 1 (Th1) or T-helper cell type 2 (Th2) cytokine production quantitated by enzyme-linked immunoabsorbent assay (ELISA). Differentiated ICOS^{-/-} cells were able to produce interferon (IFN)- γ and IL-10, but failed to express IL-4 when re-stimulated (Fig. 3a). The IL-4 defect could be rescued when the cells were differentiated in the presence of exogenous IL-4 and anti-IFN- γ antibody to promote Th2 differentiation (Fig. 3a). The differentiated ICOS^{-/-} effector cells were also defective in IL-2 production (Fig. 3a), which could not be compensated by the addition of exogenous polarizing cytokines (Fig. 3a). ICOS may therefore be essential in the regulation of IL-2 expression, not only during T-cell activation (Fig. 2c) but also in the effector phase (Fig. 3a).

To study whether the defective activation and cytokine production by ICOS-deficient cells would result in any immunodeficiency *in vivo*, we immunized wild-type or knockout mice in the footpads and base of the tail with keyhole limpet haemocyanin (KLH) protein using either alum or complete Freund's adjuvant (CFA) as adjuvant. Eight days later, the immunized mice were killed and draining lymph-node cells collected. With KLH precipitated in alum as an immunogen, ICOS^{-/-} mice had much smaller lymph nodes compared with those of the wild-type, and the number of

[†] Present address: Department of Immunology, University of Washington School of Medicine, Box 357650, Seattle, Washington 98195-7650, USA.

cells collected per node was roughly one-half of that obtained from wild-type mice (data not shown). This suggests a defect of *in vivo* activation and proliferation by ICOS^{-/-} T cells. When equal numbers of wild-type or knockout cells were re-stimulated with peptide *in vitro*, the knockout cells produced greatly reduced levels of IL-2; but their proliferation was only marginally affected (Fig. 3b). This indicates again an IL-2 production defect after *in vivo* priming in the ICOS^{-/-} mice. Thus, both the number of primed T cells, and the ability of the primed cells to produce cytokines was reduced in this system when alum was used as the adjuvant. However, when mice were immunized with CFA, which contains bacterial cell-wall products that promote stronger innate immune activation than alum, there was no significant defect in the immune response by the knockout mice in terms of *in vivo* expansion and *in vitro* proliferation after peptide re-stimulation (Fig. 3b), suggesting that strong inflammatory responses can bypass the requirement for ICOS. Notably, CD28^{-/-} mice have also been reported to have no significant defect in response to antigens in CFA¹⁰. It is unclear whether CD28 and ICOS are redundant in the response to antigen in CFA, or whether other pathways are involved.

CFA and alum also differ in the immune responses they promote:

alum induces a Th2-dominant response *in vivo*, whereas CFA induces a mixed Th1 and Th2 reaction. To examine further whether there is a defect associated with effector function by ICOS-deficient mice, we treated collected lymph-node cells from KLH-immunized mice with different doses of peptide for 4 days and measured effector cytokine production by ELISA. Little IL-4 production was observed by knockout cells from mice immunized with alum or CFA adjuvants (Fig. 3b), indicating a defect of Th2 effector function. Consistent with this, knockout mice from both immunization experiments exhibited reduced serum immunoglobulin- γ 1 (IgG1) against the KLH antigen (data not shown). In addition, when we used sheep red-blood cells as antigen, ICOS^{-/-} mice exhibited much smaller germinal centres than did wild type, and fewer IgG1-positive germinal centres (5 out of 26 germinal centres in ICOS^{-/-} mice versus 12 out of 14 in wild-type mice; to be described in detail elsewhere). As B-cell class-switching to IgG1 is dependent on Th2 cytokines, especially IL-4, these data also suggested a reduced production of IL-4 *in vivo*. When re-stimulated, the knockout cells expressed reduced levels of IL-5 (Fig. 3b), suggesting that reduced Th2 differentiation occurred *in vivo*. By contrast, the ICOS^{-/-} T cells from CFA-immunized mice secreted similar (or even

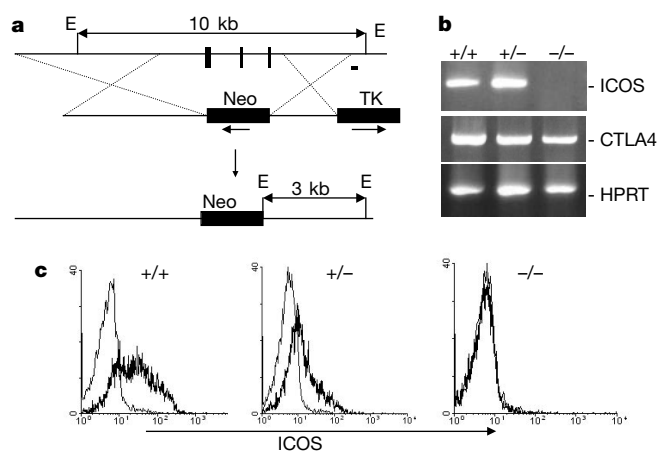


Figure 1 Generation of ICOS knockout mice. **a**, An ICOS gene-targeting construct was generated to replace three exons of the ICOS gene encoding all of the ICOS amino-acid sequences except the leader peptide with a neomycin-selecting gene driven by a PGK promoter. **b**, **c**, Lymph-node cells from ICOS^{+/+}, ICOS^{+/-} and ICOS^{-/-} mice were activated

by plate-bound anti-CD3 and anti-CD28 for 48 h and ICOS expression was assayed by RT-PCR (**b**), or staining cells with an ICOS-specific antibody (thick line; control staining is shown as thin lines) followed by flow cytometry analysis (**c**).

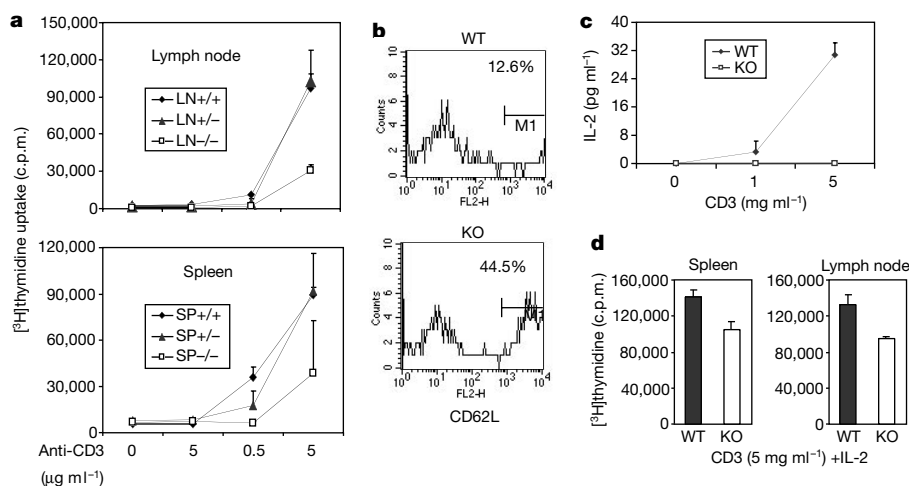


Figure 2 Defective T-cell activation *in vitro* by ICOS^{-/-} cells. **a**, **d**, Lymph-node and spleen cells were isolated from ICOS^{+/+} and ICOS^{-/-} littermates and treated in triplicates with plate-bound anti-CD3 antibody at indicated concentration at the absence (**a**) or presence (**d**) of IL-2. T-cell proliferation was assayed 3 days later by pulsing the cells with [³H]thymidine for 8 h. Data shown are from an experiment that is representative of five

independent experiments with the same result. **b**, **c**, Splenocytes from wild-type (WT) and knockout (KO) mice were treated with plate-bound anti-CD3 (5 μ g ml⁻¹). Two days later, the expression of CD62L on FACS-gated CD4 T cells was analysed by flow cytometry (**b**). IL-2 production was measured after 24 h by ELISA (**c**).

slightly higher) amounts of the Th1 cytokine IFN- γ (Fig. 3b), indicative of a normal Th1 response *in vivo*. We conclude from this experiment that ICOS is important for T-dependent immune responses *in vivo*, as it is critical for efficient T-cell priming and for production of Th2 effector cytokines, in particular IL-4.

We used a murine model to assess further the role of ICOS in immune responses. Wild-type or knockout mice were sensitized with chicken ovalbumin precipitated in alum, followed by an airway boost with the antigen. When we killed experimental mice and typed lung lavage cells, ICOS^{-/-} mice exhibited eosinophil and lymphocyte infiltration in numbers comparable to those of the wild type (Fig. 4a), suggesting that ICOS is not essential for lung inflammation in this model. When we stimulated lung lymph-node cells with ovalbumin protein, the wild-type cells produced small but detectable amounts of IL-4, whereas the knockout cells did not (Fig. 4b). Production of IL-13, another Th2 cytokine important for murine airway hyperresponsiveness, was reduced in the absence of ICOS (Fig. 4b); but wild-type and knockout cells secreted similar amounts of IL-5 (Fig. 4b) and IL-10 (data not shown). This further substantiates our earlier *in vitro* and *in vivo* data that ICOS is not required for Th2 differentiation, but rather regulates IL-4 and IL-13 production by effector cells. As the eosinophil infiltration in the lung is primarily the outcome of IL-5 expression, the data are internally consistent, although different from a report using an ICOS-immunoglobulin blocking reagent in an adoptive transfer asthma model⁹.

We further analysed the antibody responses that are more

dependent on IL-4 and IL-13 production. Sera of immunized ICOS^{-/-} mice were deficient in ovalbumin-specific IgG1 and more strikingly IgE (Fig. 4c), which supports our finding that ICOS is essential for generation of IL-4-producing cells *in vivo*. Because allergic asthma is characterized by both infiltration of inflammatory cell types (eosinophils and mast cells) in the lung and functional activation of these cells by antibodies such as IgE, our data indicate that ICOS, by regulating antibody responses, has an important role in allergic asthmatic reactions. An efficient ICOS blocker could therefore be of therapeutic importance.

We then examined the importance of ICOS in an inflammatory autoimmune disease. Experimental autoimmune encephalomyelitis (EAE) can be induced in C57BL/6 mice by injection of myelin oligodendrocyte glycoprotein (MOG) peptide 35–55 in CFA^{11,12}. In this system, CD28/B7 interaction is essential for autoimmune responses^{13–15}. When we used ICOS^{+/+} mice with the C57BL/6x129 F₂ genetic background, however, the mice only exhibited mild disease, consistent with the resistant phenotype of 129/sv mice to MOG-induced EAE (Fig. 5a). Unexpectedly, ICOS^{-/-} mice on the same genetic background uniformly developed greatly enhanced disease, and 50% of the tested mice died after disease induction (Fig. 5a). The disease exhibited by these mice was even more severe than in susceptible C57BL/6 mice (data not shown). Spinal cord and brain tissues from immunized wild-type mice exhibited only mild immune infiltration, whereas knockout mice had massive infiltrates (data not shown). In addition, flow cytometric analysis revealed more CD4⁺ T cells in the central nervous system of knockout mice, a

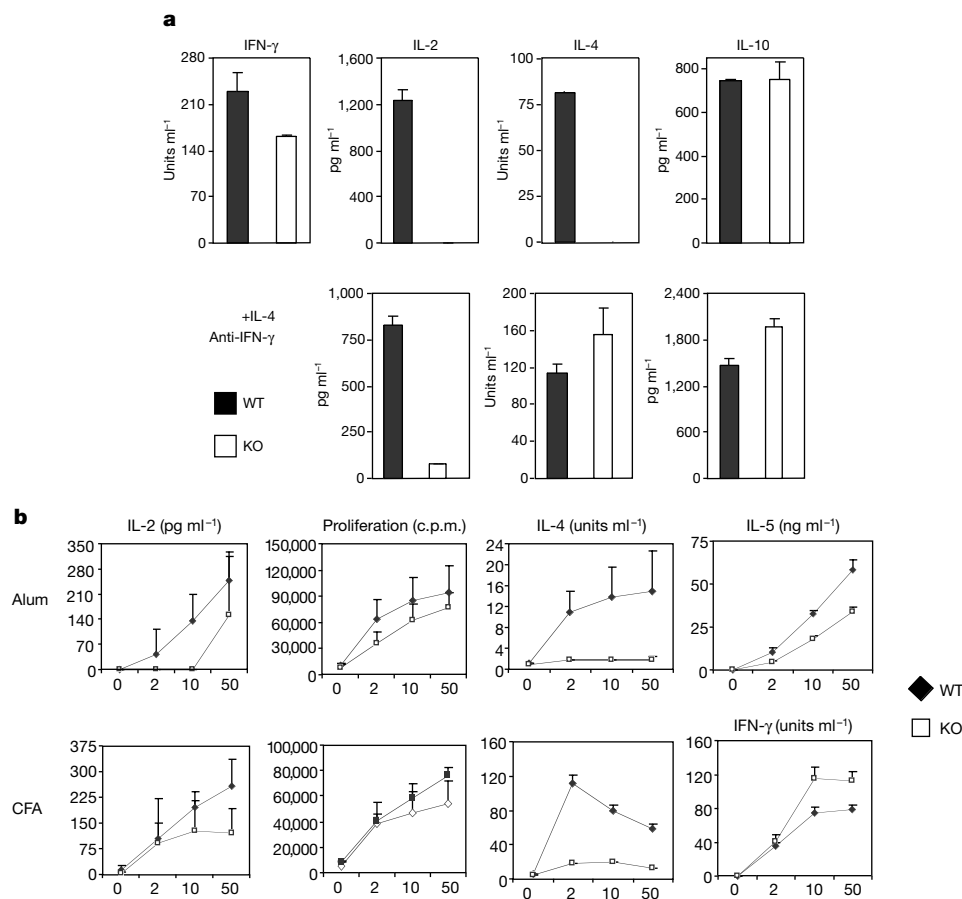


Figure 3 Defective cytokine production by ICOS^{-/-} cells. **a**, Naive wild-type (WT) or knockout (KO) T-helper cells (1×10^6 per ml) were activated with plate-bound anti-CD3, irradiated wild-type APCs (1×10^6 per ml) and IL-2 (30 units ml⁻¹) for 4 days with or without IL-4 and anti-IFN- γ . After extensive washing the differentiated cells were re-stimulated with plate-bound anti-CD3, and the cytokine production was assayed by

ELISA. **b**, Age- and sex-matched wild-type or knockout mice were immunized with KLH protein in alum or CFA adjuvant. Eight days later, draining lymph-node cells were collected and re-stimulated with KLH at indicated concentration (in μ g ml⁻¹). IL-2 production was assayed by ELISA 1 day after the treatment; proliferation, 3 days later by pulsing the cells with ³H[thymidine]; effector cytokines, 4 days later by ELISA.

significant portion of which produced IFN- γ when stimulated *in vitro* (Fig. 5c–d). These data indicate much stronger tissue inflammation in ICOS $^{-/-}$ mice.

We characterized further the differential immune responses in wild-type and knockout mice, by treating splenocytes from these mice with MOG peptide and measured T-cell cytokine production. Wild-type and knockout cells produced undetectable amounts of IL-4, but similar amounts of IL-10 and IFN- γ (data not shown). Notably, whereas the wild-type splenocytes secreted IL-13 in a dose-dependent fashion, knockout splenocytes produced no IL-13

(Fig. 5b). IL-13, as a Th2 cytokine, has an anti-inflammatory role in several systems by downregulating macrophage function^{16–19}. Furthermore, injection of recombinant IL-13 in Lewis rats inhibited EAE disease²⁰. Thus, it is likely that IL-13 deficiency in ICOS $^{-/-}$ mice rendered these mice much more susceptible to EAE, although other mechanisms may still exist.

We have examined T-cell activation and function in the absence of ICOS. ICOS $^{-/-}$ cells are defective in activation and proliferation *in vitro*. The knockout mice were also less efficient in mounting a T-cell-dependent response *in vivo*. These are probably consequences

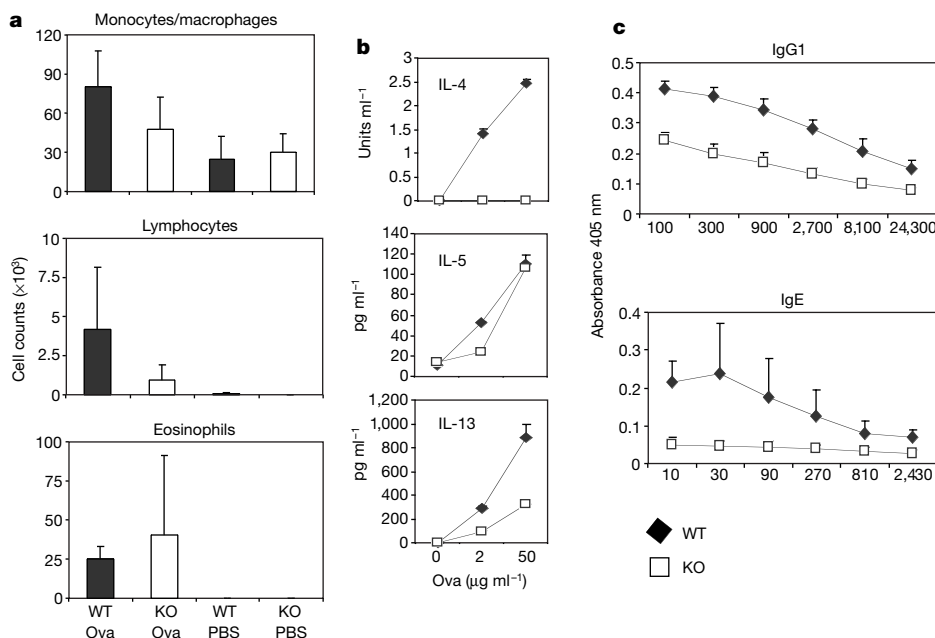


Figure 4 ICOS-deficient animals have normal infiltration of the lung with inflammatory cells but are defective in antibody responses after antigen challenges of the airway. Age-matched wild-type (WT) and knockout (KO) mice were sensitized intraperitoneally with chicken ovalbumin twice (6–7 days apart). As controls, some mice were only given PBS. Six days after the last immunization, the airway of these mice were challenged with

aerosolized antigen. **a**, The mice were killed 2 days later, and lung lavage cells were typed. **b**, Lung lymph-node cells from challenged mice were treated with the indicated protein antigen, and cytokine production was measured by ELISA. **c**, Sera from challenged mice ($n = 7$) were subjected to a threefold serial dilution, and the presence of ovalbumin-specific IgG1 and IgE was determined by ELISA (Southern Biotechnology Associate).

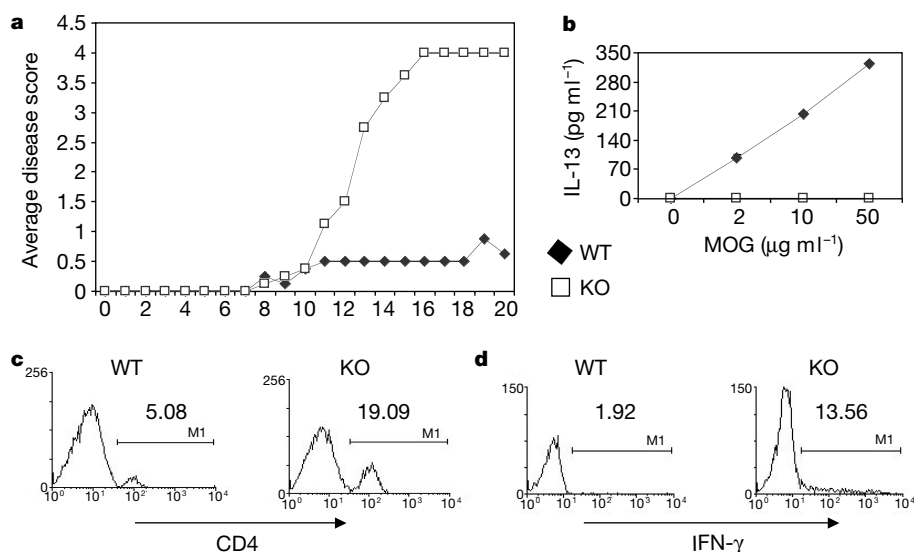


Figure 5 ICOS-deficient mice are severely susceptible to EAE. Age-matched wild-type (WT) or knockout (KO) female mice ($n = 4$) were immunized with MOG peptide 35–55 in CFA and injected with pertussis toxin according to the protocol for EAE induction. A representative of two independent experiments is shown. **a**, Time course of average disease scores in wild-type and knockout mice. The scores of two out of four ICOS-deficient mice that died are included in the graph throughout the experiment. No other

mice died. **b**, Splenocytes from immunized wild-type and knockout mice were treated with different doses of MOG peptide and IL-13 production was determined (R&D). **c**, CD4 $^{+}$ T cells were identified in mononuclear cells from CNS tissues of immunized mice by flow cytometry. **d**, These mononuclear cells were treated *in vitro* with MOG peptide and IFN- γ production by CD4 $^{+}$ T cells were assayed by intracellular cytokine staining (Pharmingen).

of a deficiency in IL-2 production, as addition of exogenous IL-2 could reconstitute proliferation *in vitro*. In addition, we also observed that *in vitro* and *in vivo* primed ICOS-deficient cells were defective in production of effector cytokines, in particular IL-4. Although ICOS appears not to be essential for Th2 differentiation in our experiments, it is critical for T-helper cell function to direct appropriate antibody responses against foreign antigens. However, in autoimmune EAE, ICOS appears have an important negative regulatory role, perhaps through the induction of IL-13, suggesting that, in contrast to the function of CD28, it may protect against inflammatory diseases. It is interesting that B7H/B7RP-1 is expressed in tissues in response to inflammatory cytokines. We propose that the induction of this ligand leads to ICOS stimulation and induction of Th2 cytokines such as IL-13, which results in the containment or resolution of the inflammatory state. In the absence of this, florid EAE results. We conclude that ICOS is a co-stimulatory receptor essential for T-cell activation and function. ICOS thus is part of a mechanism by which immunity is directed towards humoral or inflammatory responses. □

Methods

Generation of ICOS-deficient mice

We used the human ICOS amino-acid sequence to search the mouse expressed sequence tag database, and found a clone encoding the extracellular and transmembrane segment of the murine homologue. PCR primers derived from this sequence were used to amplify the mouse cDNA, which was radiolabelled to screen a genomic phage library. One phage clone was isolated, and characterization indicated that this clone contains three exons encoding all the amino acids of murine ICOS protein except the leader peptide. The ICOS gene-targeting construct, replacing three exons with a neomycin-resistance gene driven by the PGK promoter, was transfected into W9.5 ES cells and two homologous recombinants identified by Southern blot. These two clones were used to generate ICOS knockout animals. The mice used for this study are C57BL/6x129 F₂.

Production of anti-mICOS monoclonal antibody

Full-length cDNA encoding mouse ICOS (mICOS) was cloned into the expression plasmid pEGFP-N3 (Clontech) to obtain Chinese hamster ovary (CHO) cell transfectants that expressed surface mICOS–GFP (green fluorescent protein) fusion protein. To produce anti-mICOS monoclonal antibody, golden Syrian hamsters (Harland) were injected intraperitoneally with stable clones of CHO cells expressing high-level mICOS–GFP at 2-week intervals for 3 months. Five days after the final boost, hamster spleen cells were fused with the myeloma P3X63Ag8.653 (ATCC), and the resulting hybridomas screened by cell-based ELISA using Ltk– (L) cell transfectants expressing mICOS–GFP fusion protein or GFP alone. Hybridomas that reacted with L cells expressing mICOS–GFP but not with L cells expressing GFP alone were selected for further subcloning. The specificity of clone 27A12 for mICOS was confirmed by its ability to stain mICOS transfectant, but not vector or untransfected CHO cells or L cells, and for its ability to stain activated, but not unactivated T cells from C57BL/6 mice.

T-helper-cell differentiation

The methods used for isolation and treatment of peripheral CD4⁺ T cells has been described^{21,22}. Naive CD4 T cells purified from wild-type or ICOS^{−/−} mice were incubated with immobilized anti-CD3 (5 µg ml^{−1}), wild-type splenic APCs and IL-2 (a gift from Biogen) with or without IL-4 and anti-IFN-γ. Cytokine production was measured by ELISA (Pharmingen).

KLH antigen immunization

Mice (6–8 weeks old; 3 per group) were immunized with KLH (50 µg per 100 µl) emulsified in CFA or precipitated with alum in the footpads and base of the tail (100 µl per site). After 8 days, the draining lymph nodes were removed and the cells were stimulated as triplicates with or without KLH peptide. IL-2 was measured 1 day later and effector cytokines after 4 days by ELISA (Pharmingen). Cell proliferation was assayed after three days by pulsing cells with ³H[thymidine] for the last 8 h.

Airway challenge and immunoassay

Mice were sensitized and challenged essentially as described²³. Wild-type or knockout animals (3–4 mice per group) were given two intra-peritoneal injections (6–7 days apart) of 0.1 ml alum-precipitated antigen, comprising 20 µg ovalbumin (OVA; grade V, Sigma) adsorbed onto 2 mg aluminum hydroxide in PBS. Control animals received only precipitated alum in PBS. Six days after the second sensitization, mice were challenged by exposure to an aerosol of 0.5% OVA in PBS, twice for 1 h each with a 4-h interval. Two days later, mice were killed, lavage cells were flushed²⁴, and differential cell counts were performed on cytospin cell preparations stained with Diff-Quick (VWR Scientific Products, Bridgeport, NJ). Blood collected from these mice was allowed to clot at room temperature for 30 min, and serum recovered by centrifugation was analysed for anti-ovalbumin specific humoral responses by ELISA. Parathymic and peribronchial

lymph-node cells collected were stimulated (1 × 10⁶ cells per ml) for 4 days in triplicates with different doses of ovalbumin and effector cytokines measured by ELISA. The results shown represent one of two independent experiments with seven total mice in each group.

EAE induction and analysis

Murine MOG peptide 35–55 (MEVGWYRSPFSRVVHLYRNGK) was synthesized by the W. M. Keck Biotechnology Resource Center at Yale University. The peptide was purified by reverse phase (C₁₈) column high performance liquid chromatography, and a trifluoroacetic acid/acetonitrile gradient. EAE was induced by subcutaneous flank injections of 300 µg MOG 35–55 peptide in CFA (Difco, Detroit, MI) with 500 µg *Mycobacterium tuberculosis* on days 0 and 7, supplemented by intraperitoneal injections of 500 ng pertussis toxin (List Biological, Campbell, CA) as described¹². The mice were observed daily for clinical signs and scored on a scale of 0–5 with gradations of 0.5 for intermediate scores: 0, no clinical signs; 1, loss of tail tone; 2, wobbly gait; 3, hindlimb paralysis; 4, hind- and forelimb paralysis; 5, death. Preparation and stimulation of mononuclear cells from CNS tissues were done as described¹¹.

Received 30 August; accepted 20 October 2000.

- Lenschow, D. J., Walunas, T. L. & Bluestone, J. A. CD28/B7 system of T cell co-stimulation. *Annu. Rev. Immunol.* **14**, 233–258 (1996).
- Chambers, C. A. & Allison, J. P. Co-stimulatory regulation of T cell function. *Curr. Opin. Cell Biol.* **11**, 203–210 (1999).
- Thompson, C. B. & Allison, J. P. The emerging role of CTLA-4 as an immune attenuator. *Immunity* **7**, 445–450 (1997).
- Oosterwegel, M. A., Greenwald, R. J., Mandelbrot, D. A., Lorschach, R. B. & Sharpe, A. H. CTLA-4 and T cell activation. *Curr. Opin. Immunol.* **11**, 294–300 (1999).
- Hutloff, A. *et al.* ICOS is an inducible T-cell co-stimulator structurally and functionally related to CD28. *Nature* **397**, 263–266 (1999).
- Yoshinaga, S. K. *et al.* T-cell co-stimulation through B7RP-1 and ICOS. *Nature* **402**, 827–832 (1999).
- Swallow, M. M., Wallin, J. J. & Sha, W. C. B7h, a novel co-stimulatory homolog of B7.1 and B7.2, is induced by TNFα. *Immunity* **11**, 423–432 (1999).
- Kopf, M. *et al.* Inducible co-stimulator protein (ICOS) controls T helper cell subset polarization after virus and parasite infection. *J. Exp. Med.* **192**, 53–62 (2000).
- Coyle, A. J. *et al.* The CD28-related molecule ICOS is required for effective T cell–dependent immune responses. *Immunity* **13**, 95–105 (2000).
- Liu, Y. *et al.* Co-stimulation of murine CD4 T cell growth: cooperation between B7 and heat-stable antigen. *Eur. J. Immunol.* **22**, 2855–2859 (1992); erratum *ibid.* **23**, 780 (1993).
- Juedes, A. E. *et al.* Kinetics and cellular origin of cytokines in the central nervous system: insight into mechanisms of myelin oligodendrocyte glycoprotein-induced experimental autoimmune encephalomyelitis. A critical role for lymphotoxin in experimental allergic encephalomyelitis. *J. Immunol.* **164**, 419–426 (2000).
- Suen, W. E., Bergman, C. M., Hjelmstrom, P. & Ruddle, N. H. A critical role for lymphotoxin in experimental allergic encephalomyelitis. *J. Exp. Med.* **186**, 1233–1240 (1997).
- Oliveira-dos-Santos, A. J. *et al.* CD28 co-stimulation is crucial for the development of spontaneous autoimmune encephalomyelitis. *J. Immunol.* **162**, 4490–4495 (1999).
- Girvin, A. M. *et al.* A critical role for B7/CD28 co-stimulation in experimental autoimmune encephalomyelitis: a comparative study using co-stimulatory molecule-deficient mice and monoclonal antibody blockade. *J. Immunol.* **164**, 136–143 (2000).
- Chang, T. T., Jabs, C., Sobel, R. A., Kuchroo, V. K. & Sharpe, A. H. Studies in B7-deficient mice reveal a critical role for B7 co-stimulation in both induction and effector phases of experimental autoimmune encephalomyelitis. *J. Exp. Med.* **190**, 733–740 (1999).
- de Waal Malefyt, R. *et al.* Effects of IL-13 on phenotype, cytokine production, and cytotoxic function of human monocytes. Comparison with IL-4 and modulation by IFN-γ or IL-10. *J. Immunol.* **151**, 6370–6381 (1993).
- Viale, G. & Vercelli, D. Interleukin-13 regulates the phenotype and function of human monocytes. *Int. Arch. Allergy Immunol.* **107**, 176–178 (1995).
- Cosentino, G. *et al.* IL-13 down-regulates CD14 expression and TNF-α secretion in normal human monocytes. *J. Immunol.* **155**, 3145–3151 (1995).
- Kim, C. *et al.* Interleukin-13 effectively down-regulates the monocyte inflammatory potential during traumatic stress. *Arch. Surg.* **130**, 1330–1336 (1995).
- Cash, E. *et al.* Macrophage-inactivating IL-13 suppresses experimental autoimmune encephalomyelitis in rats. *J. Immunol.* **153**, 4258–4267 (1994).
- Dong, C. *et al.* JNK is required for effector T-cell function but not for T-cell activation. *Nature* **405**, 91–94 (2000).
- Dong, C. *et al.* Defective T cell differentiation in the absence of Jnk1. *Science* **282**, 2092–2095 (1998).
- Kung, T. T. *et al.* Characterization of a murine model of allergic pulmonary inflammation. *Int. Arch. Allergy Immunol.* **105**, 83–90 (1994).
- Temann, U. A., Geba, G. P., Rankin, J. A. & Flavell, R. A. Expression of interleukin 9 in the lungs of transgenic mice causes airway inflammation, mast cell hyperplasia, and bronchial hyperresponsiveness. *J. Exp. Med.* **188**, 1307–1320 (1998).

Acknowledgements

The authors thank C. L. Stewart for providing reagents; L. Evangelisti, D. Butkus, C. Hughes, J. Stein and M. Chen for technical assistance; E. Enyon for discussion; and F. Manzo for secretarial work. This work was supported by grants from the NIH (to R.A.F.), the National Multiple Sclerosis Society (to N.H.R.), the Sandler Program for Asthma Research (to R.A.F.) and the Howard Hughes Medical Institute. C.D. is a recipient of an Arthritis Investigator award, and J.A. and R.A.F. are Investigators of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to R.A.F. (e-mail: richard.flavell@yale.edu).

ICOS is critical for CD40-mediated antibody class switching

Alexander J. McAdam*, Rebecca J. Greenwald*, Michele A. Levin*, Tatyana Chernova†, Nelly Malenkovich†, Vincent Ling‡, Gordon J. Freeman†§ & Arlene H. Sharpe*§

* Departments of Pathology, Brigham and Women's Hospital and Harvard Medical School, 221 Longwood Avenue, Boston, Massachusetts 02115, USA
† Department of Adult Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA
‡ Department of Immunology, Genetics Institute, 87 CambridgePark Drive, Cambridge, Massachusetts 02081, USA
§ These authors contributed equally to this work

The inducible co-stimulatory molecule (ICOS) is a CD28 homologue implicated in regulating T-cell differentiation^{1–5}. Because co-stimulatory signals are critical for regulating T-cell activation, an understanding of co-stimulatory signals may enable the design of rational therapies for immune-mediated diseases⁶. According to the two-signal model for T-cell activation, T cells require an antigen-specific signal and a second, co-stimulatory, signal for optimal T-cell activation⁶. The co-stimulatory signal promotes T-cell proliferation, lymphokine secretion and effector function. The B7–CD28 pathway provides essential signals for T-cell activation, but does not account for all co-stimulation. We have generated mice lacking ICOS (ICOS^{−/−}) to determine the essential functions of ICOS. Here we report that ICOS^{−/−} mice exhibit profound deficits in immunoglobulin isotype class switching, accompanied by impaired germinal centre formation. Class switching was restored in ICOS^{−/−} mice by CD40 stimulation, showing that ICOS promotes T-cell/B-cell collaboration through the CD40/CD40L pathway.

ICOS expression is induced on T cells within 24 h after stimulation of the T-cell receptor, unlike CD28 which is expressed by resting T cells^{1,3,4}. The inducible expression of ICOS, and its preferential induction of interleukin (IL)-4 and IL-10 suggest that ICOS may amplify and/or regulate T-cell responses^{1,3}. ICOS binds B7h^{4,7,8} (also called B7RP-1, GL50, LICOS, B7H2; refs 8–10), which has homology to the CD28 ligands B7-1 and B7-2 (ref. 7). ICOS does not bind B7-1 or B7-2 (refs 4, 8). B7h does not bind to CD28 or CTLA-4 (ref. 8).

To investigate the role of ICOS *in vivo*, we generated mice lacking ICOS by gene targeting. The targeting vector replaced the ICOS immunoglobulin (Ig)-V-like exon with the neomycin resistance gene (Fig. 1a), eliminating B7h binding to ICOS. To confirm the absence of ICOS protein in ICOS^{−/−} mice (Fig. 1b), anti-CD3 activated CD4⁺ cells from wild-type and ICOS^{−/−} mice were stained with anti-ICOS antibody³ (Fig. 1c). Wild-type, but not ICOS^{−/−} T cells express ICOS. Further, a B7h–Ig fusion protein⁸ binds to wild-type CD4⁺ cells, but not to ICOS^{−/−} CD4⁺ cells (Fig. 1c), suggesting that ICOS is the only receptor for B7h on these cells.

ICOS^{−/−} mice are viable and born at the expected frequency. ICOS^{−/−} mice are grossly and histologically normal. Flow cytometry showed comparable numbers and maturation of thymocytes in ICOS^{−/−} and ICOS^{+/+} mice (data not shown). Spleen and lymph-node cells had comparable numbers of B220⁺, CD3⁺, CD4⁺ and CD8⁺ cells, and of the natural killer cell marker DX5 population (data not shown).

Because ICOS has been implicated in T-cell/B-cell collaboration^{2,4,11}, we investigated the role of ICOS in humoral immune responses. We first measured basal immunoglobulins. ICOS^{−/−} and ICOS^{+/+} mice had comparable IgM levels but ICOS^{−/−} mice had modestly elevated IgG3 (Fig. 2a). ICOS^{−/−} mice had reduced IgG1, IgG2a and IgE levels (Fig. 2a). IgE was unde-

tectable (< 20 ng ml^{−1}) in eight out of nine ICOS^{−/−} mice, but had a mean concentration of 170 ng ml^{−1} in ICOS^{+/+} mice. These data suggested that ICOS might have a critical role in class switching of immunoglobulin isotypes.

To examine the role of ICOS in humoral immune responses *in vivo*, we immunized ICOS^{−/−} and ICOS^{+/+} mice with a haptenated protein, 2,4,6-trinitrophenol-keyhole limpet haemocyanin (TNP-KLH), and measured anti-TNP immunoglobulin isotypes¹². We initially immunized without adjuvant, because adjuvants increase inflammatory cytokine expression, which upregulates co-stimulatory molecules including B7-1 and B7-2 (ref. 12). When immunized with TNP-KLH in saline, ICOS^{−/−} mice produced less IgG1 and IgG2a anti-TNP antibodies than did ICOS^{+/+} mice, although this did not reach statistical significance (Fig. 2b). The lack of ICOS did not affect antigen-specific IgM production, indicating that B-cell responsiveness is not directly dependent on ICOS. When ICOS^{−/−} mice were immunized with TNP-KLH in the adjuvants alum and incomplete Freund's adjuvant (IFA), marked defects in isotype class switching were observed (Fig. 2b). The anti-TNP IgM response was comparable in ICOS^{−/−} and ICOS^{+/+} mice; however, the anti-TNP IgG2a response was reduced to 6% of wild-type levels with alum or IFA in ICOS^{−/−} mice (*P* < 0.05 for both, Student's *t*-test). The anti-

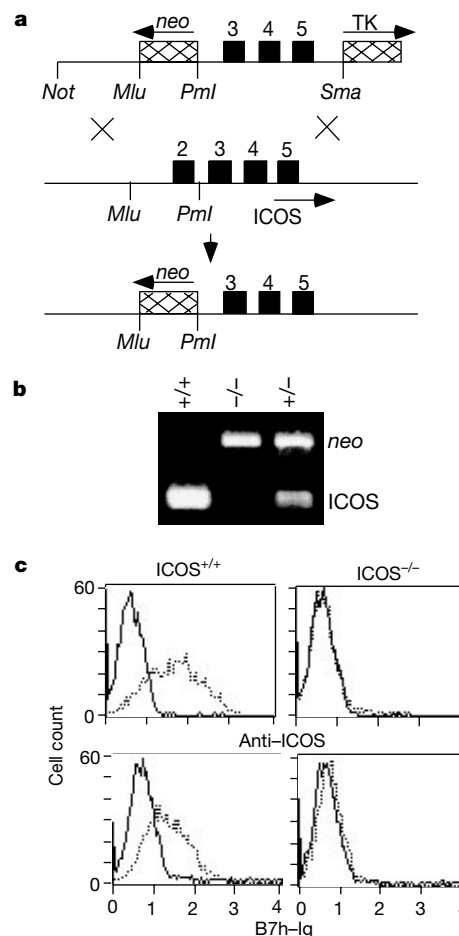


Figure 1 Generation of ICOS^{−/−} mice. **a**, The organization of the ICOS genomic locus (GenBank accession numbers AF327184 and AF327185) is shown (middle). Exons are filled boxes. In the targeting vector (top), *neo* replaced exon 2. The thymidine kinase gene (TK) lies external to the 3' genomic DNA fragment. The structure of the targeted allele is shown (bottom). **b**, DNA from ICOS^{+/+}, ICOS^{−/−} and ICOS^{+/−} mice was subjected to PCR with primer pairs to amplify regions of *neo* or exon 2 of the ICOS gene. **c**, Splenocytes from ICOS^{+/+} or ICOS^{−/−} mice were activated with anti-CD3 and stained with ICOS antibody (dotted line, top panels) or B7h–Ig (dotted line, bottom panels) or control Ig (solid lines) and CD4. The graphs are gated on CD4⁺ cells.

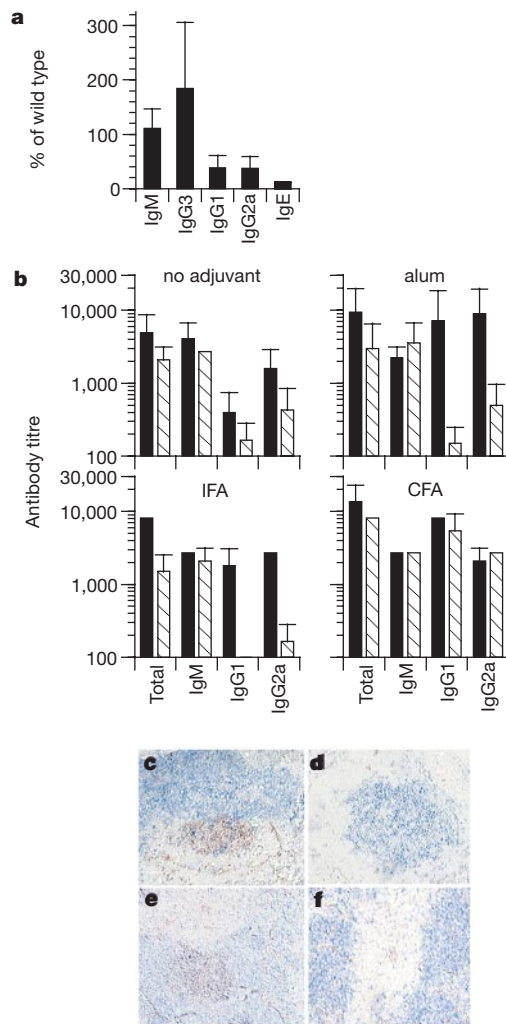


Figure 2 Antibody responses and GCs in $ICOS^{-/-}$ mice. **a**, Basal immunoglobulin levels in $ICOS^{-/-}$ mice are shown ($\pm$ one s.d.) as a percentage of $ICOS^{+/+}$ mice. **b**, $ICOS^{+/+}$ (solid bars) or $ICOS^{-/-}$ (cross-hatched bars) mice were immunized with TNP-KLH without adjuvant, in alum, IFA or CFA. The average anti-TNP titre ($\pm$ one s.d.) after 7 d is shown. **c–f**, $ICOS^{+/+}$ and $ICOS^{-/-}$ mice were immunized with TNP-KLH in IFA and 7 d later spleens were collected for immunohistochemistry. Tissue was stained for $CD4^{+}$ cells (blue stain) and GC B cells with PNA (red stain) in $ICOS^{+/+}$ (**c**) and $ICOS^{-/-}$ (**d**) mice. Tissue was stained for B220 $^{+}$ B cells (blue stain) and GC B cells with PNA (red stain) in $ICOS^{+/+}$ (**e**) and $ICOS^{-/-}$ (**f**) mice. Each panel represents the average field at $\times 100$ original magnification.

TNP IgG1 response was further reduced; $ICOS^{-/-}$ mice made 2% of wild-type levels with alum and no detectable anti-TNP IgG1 with IFA ($P < 0.05$ for both, Student's t -test). In contrast, $ICOS^{-/-}$ and $ICOS^{+/+}$ mice immunized with TNP-KLH in complete Freund's adjuvant (CFA) in the footpad, a highly inflammatory route, exhibited comparable TNP-specific antibody responses of all isotypes (Fig. 2b).

We next evaluated germinal centre (GC) formation in $ICOS^{+/+}$ and $ICOS^{-/-}$ mice¹³. At 7 d after immunization with TNP-KLH in IFA (Fig. 2c–f) or alum (data not shown), $ICOS^{-/-}$ mice (Fig. 2d, f) had smaller and fewer GCs than $ICOS^{+/+}$ mice (Fig. 2c, e). With IFA $ICOS^{-/-}$ mice had 0.24 ± 0.2 (mean $\pm$ s.d.) GCs per $\times 40$ field, whereas $ICOS^{+/+}$ mice had 6.3 ± 0.6 GCs per field. However the anatomic organization of $CD4^{+}$ and B220 $^{+}$ cells were comparable in $ICOS^{-/-}$ and $ICOS^{+/+}$ mice after immunization (Fig. 2c–f). $ICOS^{-/-}$ mice also had normal numbers of CD21 $^{+}$ follicular dendritic cells (FDCs) (data not shown) although lacking GCs. Therefore, in the absence of B7h–ICOS interactions, it is likely that B cells are poorly stimulated in the T-cell-rich regions and fail to migrate into the

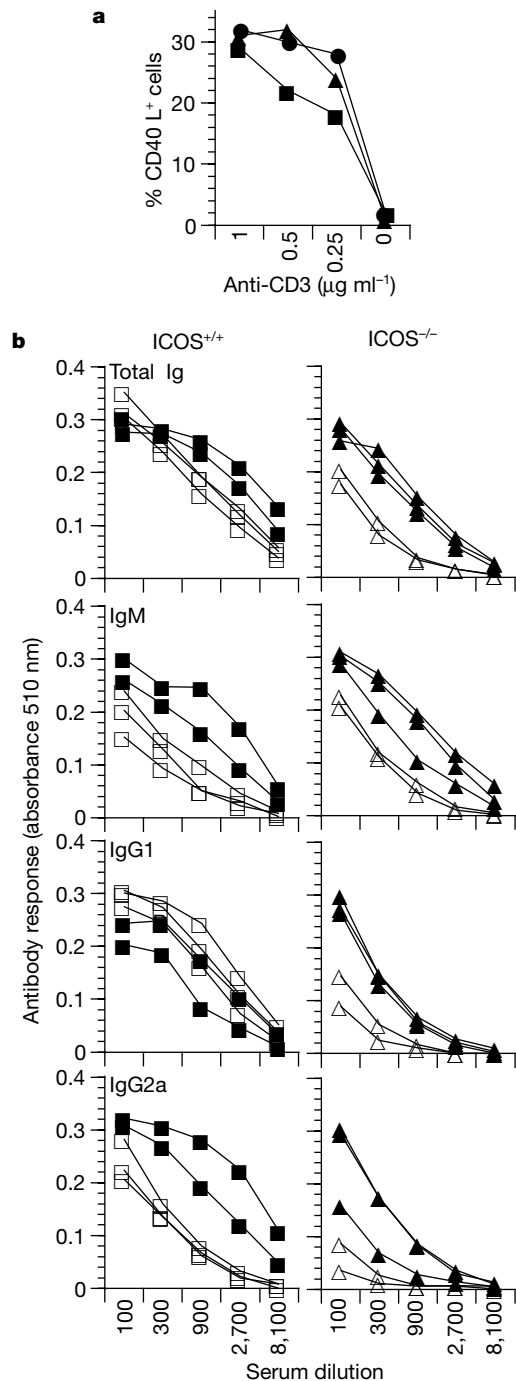


Figure 3 The role of CD40 in ICOS mediated antibody class switching. **a**, Purified wild-type $CD4^{+}$ cells were stimulated with latex beads coated with anti-CD3 and control rat Ig (squares), or anti-ICOS (circles) or anti-CD28 (triangles). After 24 h cells were stained for CD40L. **b**, $ICOS^{+/+}$ (squares) and $ICOS^{-/-}$ mice (triangles) were immunized with TNP-KLH in alum and treated with stimulatory anti-CD40 (filled symbols) or control rat IgG (open symbols) at $100 \mu\text{g d}^{-1}$, intravenously. After 7 d, mice were bled and isotype specific anti-TNP responses measured by ELISA.

FDC network to form a GC.

Stimulation of CD40 on B cells and antigen-presenting cells (APCs) is critical for class switching of T-cell-dependent antibody responses and GC formation^{14,15}. Immunohistochemical studies showed ICOS expression in the apical light zone of the GC (a site of T-cell/B-cell collaboration) in human tonsil, and *in vitro* studies showed that ICOS co-stimulation induces CD40L expression by human T cells¹. To determine whether ICOS stimulation also

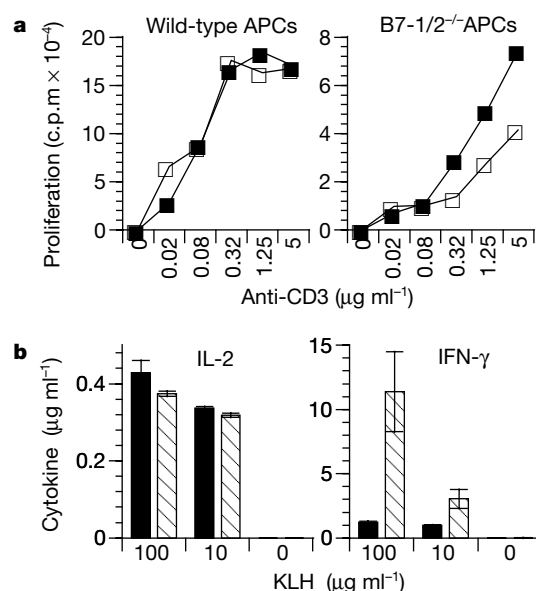


Figure 4 T-cell responses in ICOS^{-/-} mice. **a**, CD4⁺ cells from ICOS^{+/+} (filled squares) or ICOS^{-/-} (open squares) mice were stimulated with APCs from wild-type or B7-1/2^{-/-} APC and anti-CD3 for 2 d. **b**, ICOS^{+/+} (filled bars) or ICOS^{-/-} (cross-hatched bars) mice were

immunized with TNP-KLH in CFA in the footpad. CD4⁺ cells were purified from draining lymph nodes and stimulated with APCs and KLH *in vitro* and supernatants evaluated for cytokine production by ELISA. IL-2 levels are shown on day 1, and IFN- γ levels on day 4.

enhances induction of CD40L expression by mouse T cells, we stimulated CD4⁺ cells with beads coated with anti-CD3 and anti-ICOS, or anti-CD28, or control rat IgG. Stimulation of ICOS or CD28 increased the anti-CD3-stimulated CD40L induction (Fig. 3a). To ascertain whether a defect in CD40L induction caused the reduced antibody class switching in ICOS^{-/-} mice, we immunized ICOS^{+/+} and ICOS^{-/-} mice with TNP-KLH in alum, injected the mice daily with stimulatory anti-CD40 antibody¹⁶ or control IgG, and bled the mice 7 d later. CD40 stimulation restored class switching of anti-TNP IgG1 and IgG2a in ICOS^{-/-} mice (Fig. 3b). Together, these data show that there is a critical interaction between the ICOS pathway and the CD40 pathway. CD40 stimulation may directly stimulate class switching in B cells or, like CFA, enhance expression of B7-1 and B7-2 (ref. 17) to overcome the absence of ICOS.

Because antibody class switching also depends on CD4⁺ T-cell activation and cytokine production¹⁸, we investigated how ICOS affects CD4⁺ T-cell responses. When CD4⁺ cells were purified and stimulated with anti-CD3 and wild-type APCs, proliferation of ICOS^{+/+} and ICOS^{-/-} CD4⁺ cells was comparable (Fig. 4a). We next used APCs lacking B7-1 and B7-2 to determine whether ICOS stimulation could affect proliferation in the absence of these co-stimulatory molecules. Proliferation of ICOS^{-/-} CD4⁺ cells is modestly reduced compared with that of ICOS^{+/+} CD4⁺ cells on stimulation with B7-1/2^{-/-} APCs (Fig. 4a). These data show that ICOS has a small, B7-independent, co-stimulatory effect on proliferation.

We investigated the role of ICOS in cytokine production by immunizing ICOS^{-/-} and ICOS^{+/+} mice with TNP-KLH in IFA or CFA, and restimulating CD4⁺ cells from draining lymph nodes with KLH and APCs. Cells from the IFA immunized mice did not produce detectable cytokines (data not shown). CD4⁺ cells from ICOS^{-/-} and ICOS^{+/+} mice immunized with TNP-KLH in CFA produced similar levels of IL-2. Notably, CD4⁺ cells from ICOS^{-/-} mice produced more interferon (IFN)- γ than did CD4⁺ cells from ICOS^{+/+} mice (Fig. 4b). IL-4 and IL-10 were not detectable, consistent with previous studies showing that CFA primarily enhances a T-helper-cell type 1 (Th1) response¹⁹. These data indicate that the absence of ICOS may increase generation of Th1 cells, and is consistent with previous studies suggesting that ICOS stimulation allows enhanced Th2 differentiation¹⁻³.

The ICOS^{-/-} mouse shows that ICOS has a critical role in immunoglobulin class switching and GC formation. This deficit can be overcome by CD40 stimulation, suggesting that ICOS promotes isotype switching by regulating the CD40-CD40L pathway. The class-switching defect observed in ICOS^{-/-} mice is similar to that in B7-1/2^{-/-} mice, but differs in that B7-1/2^{-/-} mice are unable to class switch or form GCs irrespective of the immunization route or adjuvant used¹². These differences suggest that the strong inflammatory response elicited by CFA immunization in the footpad induces expression of B7-1 or B7-2, eliciting class switching the absence of ICOS. Therefore, there are overlapping and distinct critical roles for ICOS and B7-CD28 interactions in T-cell/B-cell collaboration. The role of ICOS in antibody responses suggests that manipulation of this pathway may have therapeutic benefits. Stimulation of ICOS might enhance the efficacy of vaccines, whereas inhibition of the ICOS pathway might be used to treat diseases with pathological IgE production, such as asthma, allergies or auto-antibody production. □

Methods

ICOS detection

Chimaeric mice generated from targeted embryonic stem cells had heterozygous progeny, which were interbred to generate ICOS^{-/-} mice. Genotyping was performed by polymerization chain reaction (PCR) as described¹². The sequences of *neo* primers, which amplify a 479-base-pair (bp) product, were 5'-ATTGAACAAGATGGATTGCAC-3' and 5'-TCTTCGTCAGATCATCCT-3', and sequences of ICOS primers, which amplify a 298-bp region of exon 2 deleted in ICOS^{-/-} mice, were 5'-GGTTTGTGTTGCTTGGCTATA-3' and 5'-GGTCTTGGTGAGTTCGCAGAG-3'. Appropriate integration of the targeting vector was confirmed by Southern analysis¹². The genomic probe used for Southern was external to genomic DNA fragments used in the targeting vector. *SacI* digest yields a 15-kb band in wild-type DNA, and a 9.4-kb band in the targeted allele.

To detect expression of ICOS, 2×10^6 splenocytes from ICOS^{+/+} or ICOS^{-/-} mice were activated for 24 h with anti-CD3 (145-2C11, 1 $\mu\text{g ml}^{-1}$) in 2 ml of media. A monoclonal antibody against ICOS³ conjugated to FITC or a B7h-Ig⁸ fusion protein conjugated to FITC and anti-CD4-PE (Pharmingen) were used to stain the splenocytes for analysis by flow cytometry.

Antibody responses and basal Ig detection

Mice were immunized with 100 μg of TNP-KLH alum (Pierce) intraperitoneally, IFA (Sigma) in the footpad, or CFA (containing killed mycobacteria) in the footpad (Sigma)¹². For studies using anti-CD40, ICOS^{-/-} mice were treated with antibody against CD40 (ref. 16) or rat IgG (Sigma), 100 μg each day intravenously, for a week starting on the day of immunization (anti-CD40 was 3/23, provided by G. Klaus). Mice were bled 7 and 14 d later, and the sera were tested for isotype specific anti-TNP antibodies by enzyme-linked

immunoassay (ELISA) as described¹². The endpoint titre for each mouse was calculated as the greatest dilution of serum giving a signal (optical density) of 10% of the maximum signal within that ELISA²⁰. Basal immunoglobulin levels were determined by ELISA using isotype-specific goat antibodies to mouse immunoglobulin (Southern). Purified mouse Ig standards (from Southern, except IgG3 and IgE, which were from Pharmingen) were used to quantify antibodies in mouse serum.

Immunohistochemistry

ICOS^{+/+} and ICOS^{-/-} mice were immunized with TNP-KLH in IFA and at day 7, spleens were collected, frozen, and 5-µm thick sections were cut. Tissue sections were stained as described²¹ for CD4 (Pharmingen) and B220 (Pharmingen) using biotinylated antibodies followed by streptavidin-alkaline phosphatase, and GC B cells with peanut agglutinin/horseradish-peroxidase (Sigma).

T-cell responses

To measure the effect of ICOS stimulation on CD40L induction, 10⁵ CD4⁺ cells purified by magnetic cell sorting (Millitene) were stimulated with 10⁵ latex beads (Interfacial Dynamics). Beads were conjugated to anti-CD3 (0–1 µg ml⁻¹) and either anti-ICOS, anti-CD28 or control rat IgG (3 µg ml⁻¹). After 24 h, cells were stained with anti-CD40L-PE and anti-CD4-FITC (Pharmingen) for analysis by flow cytometry.

To assay proliferation, 5 × 10⁵ purified CD4⁺ cells from ICOS^{+/+} or ICOS^{-/-} were stimulated with 3 × 10⁵ T-cell-depleted mitomycin-C-treated APCs from wild-type or B7-1/2^{-/-} 129/SvJae mice¹² with titrated anti-CD3 in 96-well flat-bottom plates. After 2 d, the cells were pulsed with [³H]thymidine for 6 h. To measure cytokine production, CD4⁺ cells were purified from the draining lymph nodes of ICOS^{+/+} or ICOS^{-/-} mice immunized 10 d earlier with TNP-KLH in CFA in the footpad. CD4⁺ cells (2 × 10⁵) were cultured with APCs (3 × 10⁵) from wild-type 129/SvJae mice and KLH in 24-well plates. Supernatants were assayed for cytokines by ELISA²².

Received 12 September; accepted 7 November 2000.

- McAdam, A. J., Schweitzer, A. N. & Sharpe, A. H. The role of B7 co-stimulation in activation and differentiation of CD4⁺ and CD8⁺ T cells. *Immunol. Rev.* **165**, 231–247 (1998).
- Hutloff, A. *et al.* ICOS is an inducible T-cell co-stimulator structurally and functionally related to CD28. *Nature* **397**, 263–266 (1999).
- Coyle, A. J. *et al.* The CD28-related molecule ICOS is required for effective T cell-dependent immune responses. *Immunity* **13**, 95–105 (2000).
- McAdam, A. J. *et al.* Mouse inducible costimulatory molecule (ICOS) expression is enhanced by CD28 costimulation and regulates differentiation of CD4⁺ T cells. *J. Immunol.* **165**, 5035–5040 (2000).
- Yoshinaga, S. K. *et al.* T-cell co-stimulation through B7RP-1 and ICOS. *Nature* **402**, 827–832 (1999).
- Mages, H. W. *et al.* Molecular cloning and characterization of murine ICOS and identification of B7h as ICOS ligand. *Eur. J. Immunol.* **30**, 1040–1047 (2000).
- Swallow, M. M., Wallin, J. J. & Sha, W. C. B7h, a novel costimulatory homolog of B7.1 and B7.2, is induced by TNFα. *Immunity* **11**, 423–432 (1999).
- Ling, V. *et al.* Cutting edge: identification of GL50, a novel B7-like protein that functionally binds to ICOS receptor. *J. Immunol.* **164**, 1653–1657 (2000).
- Brodie, D. *et al.* ICOS, a primordial costimulatory ligand? *Curr. Biol.* **10**, 333–336 (2000).
- Wang, S. *et al.* Costimulation of T cells by B7-H2, a B7-like molecule that binds ICOS. *Blood* **96**, 2808–2813 (2000).
- Kopf, M. *et al.* Inducible costimulator protein (ICOS) controls T helper cell subset polarization after virus and parasite infection. *J. Exp. Med.* **192**, 53–61 (2000).
- Borriello, F. *et al.* B7-1 and B7-2 have overlapping, critical roles in immunoglobulin class switching and germinal center formation. *Immunity* **6**, 303–313 (1997).
- Tarlington, D. Germinal centers: form and function. *Curr. Opin. Immunol.* **10**, 245–251 (1998).
- Kawabe, T. *et al.* The immune responses in CD40-deficient mice: impaired immunoglobulin class switching and germinal center formation. *Immunity* **1**, 167–178 (1994).
- Borrow, P. *et al.* CD40L-deficient mice show deficits in antiviral immunity and have an impaired memory CD8⁺ CTL response. *J. Exp. Med.* **183**, 2129–2142 (1996).
- Hasbold, J., Johnson-Leger, C., Atkins, C. J., Clark, E. A. & Klaus, G. G. Properties of mouse CD40: cellular distribution of CD40 and B cell activation by monoclonal anti-mouse CD40 antibodies. *Eur. J. Immunol.* **24**, 1835–1842 (1994).
- Roy, M. *et al.* Studies on the interdependence of gp39 and B7 expression and function during antigen-specific immune responses. *Eur. J. Immunol.* **25**, 596–603 (1995).
- Mosmann, T. R. & Sad, S. The expanding universe of T-cell subsets: Th1, Th2 and more. *Immunol. Today* **17**, 138–146 (1996).
- Yip, H. C. *et al.* Adjuvant-guided type-1 and type-2 immunity: infectious/noninfectious dichotomy defines the class of response. *J. Immunol.* **162**, 3942–3949 (1999).
- McAdam, A. J., Farkash, E. A., Gewurz, B. E. & Sharpe, A. H. B7 costimulation is critical for antibody class switching and CD8⁺ cytotoxic T-lymphocyte generation in the host response to vesicular stomatitis virus. *J. Virol.* **74**, 203–208 (2000).
- Greenwald, R. J. *et al.* Effects of blocking B7-1 and B7-2 interactions during a type 2 *in vivo* immune response. *J. Immunol.* **158**, 4088–4096 (1997).
- Schweitzer, A. N., Borriello, F., Wong, R. C., Abbas, A. K. & Sharpe, A. H. Role of costimulators in T cell differentiation: studies using antigen-presenting cells lacking expression of CD80 or CD86. *J. Immunol.* **158**, 2713–2722 (1997).

Acknowledgements

This work was supported by grants from Genetics Institute (to A.H.S.) and the NIH (to A.J.M., R.J.G., G.J.F. and A.H.S.). We thank M. Collins for thoughtful discussions, and L. Du, J. Burgess and B. Chang for technical support.

Correspondence and requests for materials should be addressed to A.H.S. (e-mail: asharpe@rics.bwh.harvard.edu).

ICOS is essential for effective T-helper-cell responses

Anna Tafuri^{†‡‡}, Arda Shahinian^{†‡‡}, Friedhelm Bladt^{†§}, Steve K. Yoshinaga^{||}, Manel Jordana[¶], Andrew Wakeham^{†‡‡}, Louis-Martin Boucher^{†‡‡}, Denis Bouchard^{†‡‡}, Vera S. F. Chan^{‡‡}, Gordon Duncan^{†‡‡}, Bernhard Odermatt[‡], Alexandra Ho^{†‡‡}, Annick Itie^{†‡‡}, Tom Horan^{||}, John S. Whoriskey^{||}, Tony Pawson[§], Josef M. Penninger^{†‡‡}, Pamela S. Ohashi^{‡‡} & Tak W. Mak^{†‡‡}

[†] Amgen Institute, 620 University Avenue, Toronto, Ontario M5G 2C1, Canada

[‡] Ontario Cancer Institute, and Departments of Medical Biophysics and Immunology, University of Toronto, Toronto, Ontario M5G 2C1, Canada

[§] Mount Sinai Hospital, Samuel Lunenfeld Research Institute, 600 University Avenue, Toronto, Ontario M5G 1X5, Canada

^{||} Amgen, One Amgen Center Drive, Thousand Oaks, California 91320-1789, USA

[¶] Department of Pathology and Molecular Medicine, Faculty of Health Science, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

^{*} University Hospital Zurich, Department of Pathology, CH-8091 Zurich, Switzerland

* These authors contributed equally to this work

The outcome of T-cell responses after T-cell encounter with specific antigens is modulated by co-stimulatory signals, which are required for both lymphocyte activation and development of adaptive immunity^{1–3}. ICOS^{4,5}, an inducible co-stimulator with homology to CD28, is expressed on activated, but not resting T cells, and shows T-cell co-stimulatory function *in vitro*. ICOS binds specifically to its counter-receptor B7RP-1 (refs 5–7), but not to B7-1 or B7-2. Here we provide *in vivo* genetic evidence that ICOS delivers a co-stimulatory signal that is essential both for efficient interaction between T and B cells and for normal antibody responses to T-cell-dependent antigens. To determine the physiological function of ICOS, we generated and characterized gene-targeted ICOS-deficient mice. *In vivo*, a lack of ICOS results in severely deficient T-cell-dependent B-cell responses. Germinal centre formation is impaired and immunoglobulin class switching, including production of allergy-mediating IgE, is defective. ICOS-deficient T cells primed *in vivo* and restimulated *in vitro* with specific antigen produce only low levels of interleukin-4, but remain fully competent to produce interferon-γ.

To determine whether ICOS has an integral function in T-cell co-stimulation *in vivo*, and to identify physiological situations that require ICOS activity, we disrupted the ICOS gene in mice by deleting a genomic fragment corresponding to residues 318–591 of the complementary DNA sequence⁵ (Fig. 1a). We confirmed germline transmission of the ICOS mutation by Southern blotting (Fig. 1b). ICOS^{-/-} mice generated by the intercrossing of heterozygous offspring were born at the expected mendelian frequency and were viable, fertile and of normal size. To confirm that the ICOS mutation had abolished ICOS expression, we analysed activated T cells from ICOS^{-/-} mice and control littermates by flow cytometry. ICOS was expressed on the surface of both CD4⁺ and CD8⁺ wild-type T cells, but was undetectable on ICOS^{-/-} T cells (Fig. 1c).

We next assessed the phenotypes of developing and mature T and B cells in the central and peripheral lymphoid compartments of ICOS^{-/-} mice using flow cytometry. Comparison between wild-type and ICOS^{-/-} thymus and bone marrow did not show any obvious changes in the number or ratio of cells at different maturational stages (data not shown). The expression of CD3, TCRαβ, CD28 and CD44 on mature T cells, and B220, immunoglobulin (Ig)M and IgD on mature B cells, was equal on cells isolated from spleen and lymph nodes of ICOS^{-/-} and wild-type mice (data not shown). Thus, a lack of ICOS does not result in gross alterations of T- and B-cell development.

We then examined the ability of ICOS-deficient T cells to proliferate in response to various mitogenic stimuli. The proliferation of ICOS^{-/-} T cells stimulated with a B7RP1-immunoglobulin fusion protein (B7RP1-Fc)⁵ in combination with anti-CD3 monoclonal antibody was profoundly impaired compared with wild-type T cells in the early stages of proliferation (24 h); however, recovery occurred at later time points (Fig. 1d). These data indicate a possible selective effect of ICOS co-stimulation on the early phases of T-cell proliferation and/or a later compensatory effect by other molecules. The reduced proliferation of ICOS^{-/-} T cells was associated with a markedly decreased expression of the T-cell activation markers CD40 ligand (CD40L), CD25 and CD69 (Fig. 1e). Notably, ICOS^{-/-} T cells stimulated with anti-CD3 plus anti-CD28 monoclonal antibody, concanavalin A (ConA), or phorbol myristate

acetate (PMA) plus ionomycin (PMA/Iono) proliferated normally (data not shown). ICOS^{-/-} B cells stimulated with bacterial lipopolysaccharide, anti-CD40 or anti-IgM also showed normal proliferation (data not shown). These findings indicate that ICOS deficiency may selectively abrogate T-cell responses that require ICOS-mediated co-stimulation, but does not affect, at least *in vitro*, the integrity of the CD28-B7 pathway or the intracellular machinery responsible for T- and B-cell proliferation.

ICOS may be involved in the interaction between CD4⁺ T-helper cells and B cells^{5,6,8}. To test this hypothesis, we determined the basal serum levels of IgM, IgG1, IgG2a, IgG2b, IgG3 and IgA in 8-week-old ICOS^{-/-} and littermate control mice (Fig. 2a-f). We assessed statistical significance using a test⁹ that examines trends in related groups. The basal serum levels of IgG1 were reduced significantly in

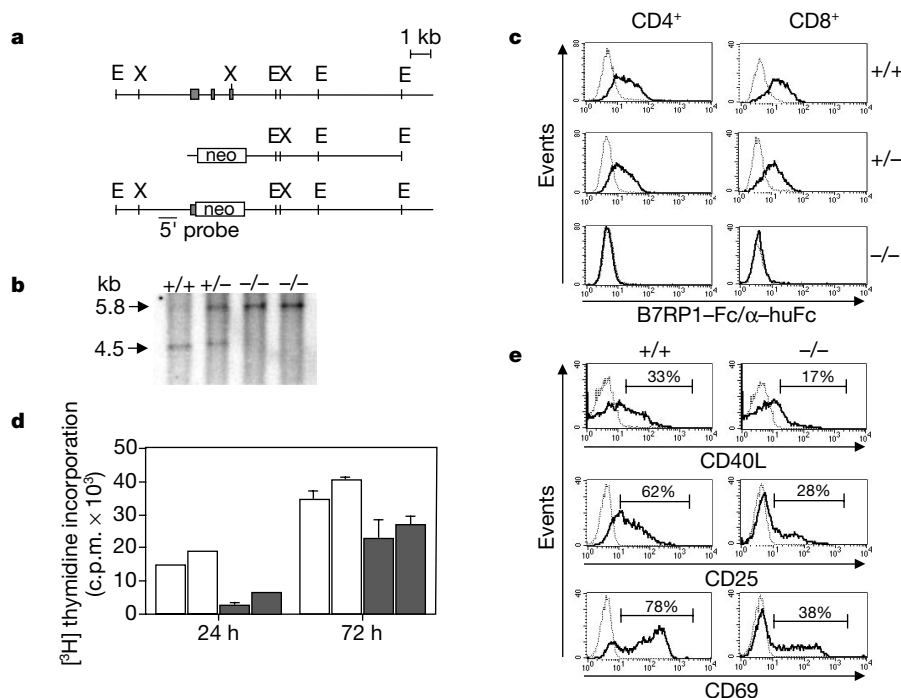


Figure 1 Generation of ICOS-deficient mice. **a**, Gene targeting in embryonic stem cells. Top, a portion of the genomic mouse *ICOS* locus showing relevant restriction sites. E, *EcoRI*; X, *XbaI*. Exons are shown as grey boxes. The structure of the pTVneo poly(A) targeting vector (middle), and the targeted *ICOS* locus (bottom) are shown. The short-arm hybridization probe used for Southern analysis is indicated. **b**, Southern blot analysis of tail biopsies from ICOS^{-/-} mice. Genomic DNA was digested with *XbaI* and hybridized with the short-arm flanking probe indicated in **a**. **c**, Lack of ICOS expression on activated ICOS^{-/-} T cells (see Methods: 'In vitro Stimulations'). Expression of ICOS (B7RP1 ligand; solid line) and background staining with an irrelevant Fc chimeric protein (dashed line) are shown

on gated T cells. Activated T cells from ICOS^{+/-} mice show lower levels of ICOS expression (mean fluorescence, MF = 51.9) as compared with wild-type (WT) cells (MF = 73.7), suggesting a gene dosage effect in heterozygous animals. **d**, Impaired proliferative response of ICOS^{-/-} T cells. Enriched T cells from ICOS^{-/-} (black bars; *n* = 2) and WT (white bars; *n* = 2) littermate mice were stimulated with anti-CD3 and B7RP1-Fc. Data are representative of one of three independent experiments. **e**, Reduced expression of activation markers on activated ICOS^{-/-} T cells. ICOS^{-/-} T cells were stimulated with anti-CD3/B7RP1-Fc and assessed 24 h later by flow cytometric analysis gated on high forward side scatter CD4⁺ T cells.

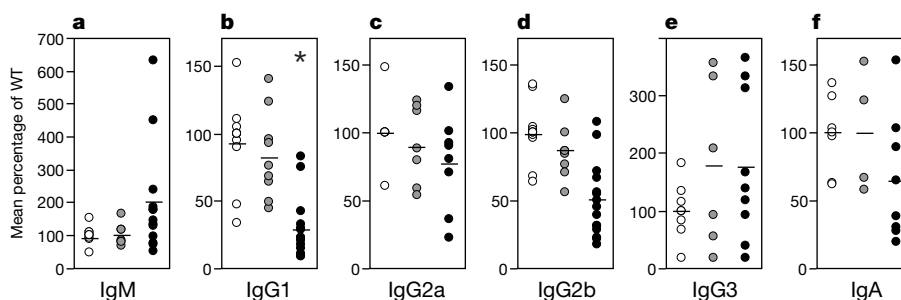


Figure 2 Reduced basal IgG1 levels in ICOS^{-/-} mice. Sera from 8-week-old WT (clear circles), ICOS^{+/-} (grey circles) and ICOS^{-/-} (black circles) mice were analysed by ELISA for levels of IgM (**a**), IgG1 (**b**), IgG2a (**c**), IgG2b (**d**), IgG3 (**e**) and IgA (**f**). Titres of WT controls were averaged in each litter and taken as 100%, and the antibody titre of each mouse was

calculated as a percentage of the mean of WT. The mean antibody titre for each group is indicated (horizontal line). IgG1 levels in ICOS^{-/-} mice are significantly lower than those in control littermates (*P* < 0.0001). Data are from three independent experiments. Asterisk indicates statistically significant results.

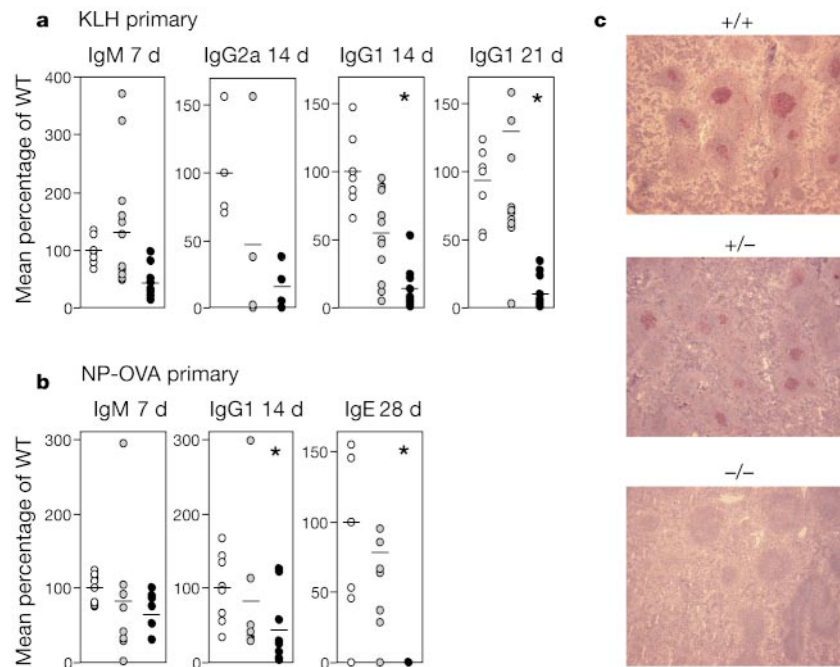


Figure 3 Impaired primary antibody responses and germinal centre formation in $ICOS^{-/-}$ mice challenged with T-cell-dependent antigens. **a**, KLH-specific primary antibody responses. WT (clear circles), $ICOS^{+/-}$ (grey circles) and $ICOS^{-/-}$ (black circles) mice were immunized with KLH in CFA. IgG1 levels in $ICOS^{-/-}$ mice were significantly reduced compared with controls at both day 14 ($P < 0.0001$) and day 21 ($P = 0.0001$). **b**, NP-OVA-specific primary antibody responses. WT, $ICOS^{+/-}$ and $ICOS^{-/-}$ mice were immunized with NP-OVA adsorbed to alum and sera were collected on days 7, 14 and 28 after immunization (labelling as in **a**). IgG1 was significantly ($P = 0.0039$) lower in $ICOS^{-/-}$

mice than in control littermates. OVA-specific IgE was undetectable in all 8 $ICOS^{-/-}$ mice tested ($P = 0.0033$), but readily detected in 5/7 WT and 7/8 $ICOS^{+/-}$ littermates. **c**, Impaired germinal centre formation in $ICOS^{-/-}$ mice in response to KLH/CFA. Spleens were snap frozen and germinal centre formation was analysed by staining frozen sections with PNA. $ICOS^{-/-}$ mice (bottom) show a marked reduction in germinal centre formation compared with WT (top) and heterozygous (centre) littermate controls. Levels of KLH-specific IgM, IgG2a and IgG1 (**a**), and NP-specific IgM, NP-specific IgG1 and OVA-specific IgE (**b**) were assessed by ELISA. Asterisks indicate statistically significant results.

$ICOS^{-/-}$ mice (30% of wild type), whereas the levels of the other immunoglobulin isotypes did not differ significantly from those of control littermates.

Basal levels of IgG1 are highly dependent on T-helper cells, as shown by their selective reduction in major histocompatibility complex class-II-deficient mice¹⁰. We therefore investigated the role of ICOS in humoral responses to T-cell-dependent antigens. $ICOS^{-/-}$ and littermate control mice were immunized with keyhole limpet haemocyanin (KLH) in complete Freund's adjuvant (CFA). As shown in Fig. 3a, titres of KLH-specific IgG1 and IgG2a were reduced markedly in the serum of $ICOS^{-/-}$ mice as compared with controls at 14 d post-challenge (both 16% of wild type), with a further reduction in IgG1 at 21 d post-challenge (10% of wild type). In response to a second T-cell-dependent antigen, nitrophenol-conjugated ovalbumin (NP-OVA) adsorbed to alum, $ICOS^{-/-}$ mice produced lower levels (43% of wild type) of NP-specific IgG1 compared with controls (Fig. 3b). Notably, OVA-specific IgE was undetectable in all $ICOS^{-/-}$ mice examined, whereas isotype switching to IgE occurred in 72% of wild-type and 87.5% of heterozygous littermates.

During T-cell-dependent B-cell responses, the formation of germinal centres is the anatomical hallmark of T-helper-cell activity and the physical site of T-cell/B-cell interaction¹¹. To investigate the involvement of ICOS in germinal centre formation, spleens of $ICOS^{-/-}$ mice and littermate controls immunized with KLH were analysed by immunohistochemistry. Staining with peanut agglutinin (PNA¹²; Fig. 3c), anti-B220, or anti-CD4 monoclonal antibody (data not shown) revealed significant reductions in the number (wild type, 36.5 ± 2.1 ; $ICOS^{+/-}$, 25.5 ± 5 ; $ICOS^{-/-}$, 4.6 ± 4.6) and size (greater than fourfold) of germinal centres in $ICOS^{-/-}$ mice compared with wild-type controls. Similar results were obtained in $ICOS^{-/-}$ mice immunized with NP-OVA (data not shown). These findings provide an anatomical basis for the impaired isotype

switching observed in $ICOS^{-/-}$ mice and suggest that ICOS is actively involved in promoting germinal centre formation and effective interaction between T and B cells.

We next investigated whether antigenic boosting with T-cell-dependent antigens could restore normal isotype switching in $ICOS^{-/-}$ mice. On secondary stimulation with KLH, the levels of IgG1 and IgG2a produced in $ICOS^{-/-}$ mice were reduced markedly (9% and 19% of wild type, respectively) compared with those in control littermates (Fig. 4a). Moreover, tertiary stimulation of $ICOS^{-/-}$ mice with NP-OVA resulted in barely detectable IgE levels that averaged only 12.8% of wild-type IgE levels (Fig. 4b). Thus, ICOS is required for isotype switching in both primary and secondary T-cell-dependent B-cell-responses.

If ICOS is required for efficient T-cell help, then B-cell responses to T-cell-independent antigens should remain unaltered in the absence of ICOS. We challenged $ICOS^{-/-}$ mice with trinitrophenol conjugated to Ficoll (TNP-Ficoll), a classical T-cell-independent antigen¹³. Isotype switching to IgG3, which in this case is T-cell-independent, occurred to the same extent in $ICOS^{-/-}$ and control mice (Fig. 3c). Thus, B cells are intrinsically normal in $ICOS^{-/-}$ mice, and the observed defects in isotype switching are probably due to alterations in T-helper-cell activity.

Interferon (IFN)- γ and interleukin (IL)-4 modulate B-cell isotype switching to IgG2a or IgG1 and IgE¹⁴⁻¹⁶, respectively. To investigate the involvement of ICOS in skewing T-cell polarization towards IFN- γ -producing T-helper-cell type 1 (Th1) versus IL-4-producing Th2 effector cells^{17,18}, CD4⁺ T cells from $ICOS^{-/-}$ and wild-type littermate controls immunized *in vivo* were re-stimulated *in vitro* with KLH. We assessed production of IL-4 and IFN- γ in culture supernatants by enzyme-linked immunosorbent assay (ELISA), and at the intracellular level by cytofluorimetry. IL-4 was profoundly reduced in the culture supernatant of $ICOS^{-/-}$ T cells

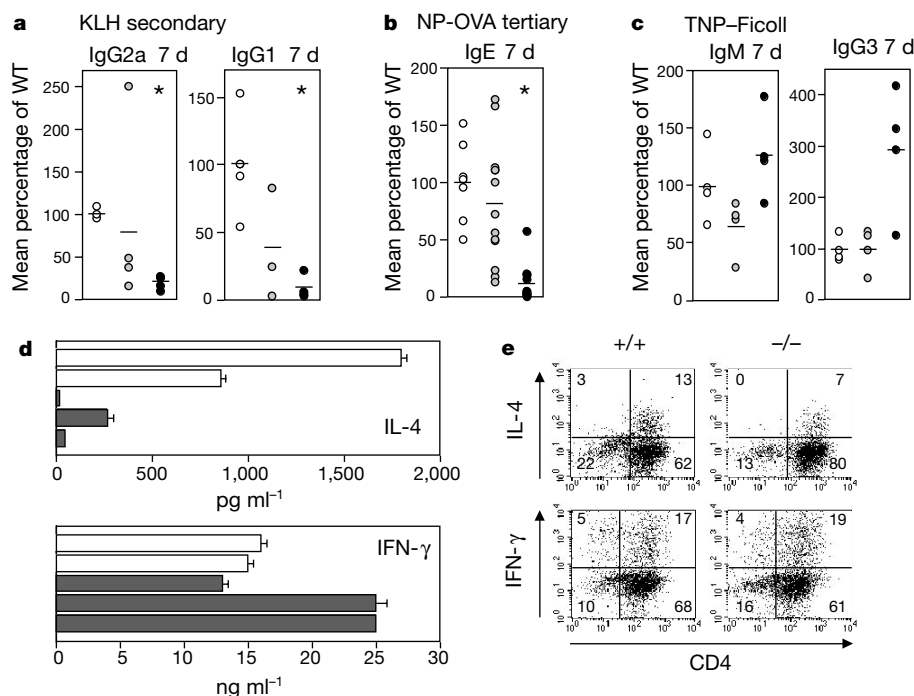


Figure 4 Impaired secondary antibody responses to T-cell-dependent antigens but normal responses to T-independent antigens in $ICOS^{-/-}$ mice. **a**, **b**, Antigenic boosting does not correct the isotype switching defect of $ICOS^{-/-}$ mice. **a**, Secondary challenge of KLH-primed mice. The levels of both IgG1 ($P = 0.0030$) and IgG2a ($P = 0.0128$) are significantly lower in $ICOS^{-/-}$ mice than in control littermates. Titres of WT controls are averaged in each litter and taken as 100% (**a–c**). Individual monoclonal antibody titres were calculated as percentage of this value. Data are from three independent experiments. **b**, Levels of OVA-specific IgE in mice immunized with NP-OVA after tertiary antigenic challenge. IgE levels are significantly lower ($P = 0.0001$) in $ICOS^{-/-}$ mice than in controls. **c**, ICOS is not required for antibody responses to T-cell-independent antigens. Sera from WT (clear circles), $ICOS^{+/+}$ (grey circles) and $ICOS^{-/-}$ (black circles) mice

compared with the wild type, but IFN- γ levels were similar (Fig. 4d). Intracellular staining revealed a reduction in the percentage of IL-4-producing cells in $ICOS^{-/-}$ T-cell cultures ($5.6 \pm 1.0\%$; $n = 5$) as compared with wild-type controls ($10.8 \pm 2.6\%$; $n = 4$) (Fig. 4e), whereas the percentage of IFN- γ -producing cells was similar (wild type, $15.3 \pm 6.5\%$; $n = 7$ versus $ICOS^{-/-}$, $16.7 \pm 4.9\%$; $n = 8$).

The deficiency in IL-4 production in the absence of ICOS may account for the defect in isotype switching to IgE and IgG1 (ref. 19). However, mechanisms dependent on cell contact that are mediated directly by ICOS or through upregulation or stabilization of surface expression of other molecules, such as CD40L, may also be involved. Although our findings support a role for ICOS in promoting Th2 differentiation⁸, it remains to be determined whether ICOS is required *in vivo* for Th2 polarization in response to infection. *In vivo* treatment with ICOS-immunoglobulin inhibits the generation of both Th1 and Th2 responses to various infectious agents²⁰. ICOS may also be a potential target for pharmacological intervention in disorders such as allergy and asthma, as a severe reduction in IgE was observed in ICOS-deficient mice even after repeated antigenic challenges. Finally, our findings suggest that ICOS can be ranked with CD28 and CD40L in the co-stimulatory pantheon. All three molecules are required for germinal centre formation^{21–23} and optimal isotype switching^{22–24} and have pivotal and non-redundant functions in regulating the crosstalk between T cells, antigen-presenting cells and B cells^{1,2,25}. Discerning the precise functions of each molecule in pathological conditions such as atopy, autoimmune diseases, and transplant and tumour rejection may be of crucial importance in identifying better tools for the treatment of these disorders. □

immunized with TNP-Ficoll were tested on day 7 after immunization for levels of TNP-specific IgM and IgG3. No significant differences in TNP-specific immunoglobulin production were observed. **d**, Impaired production of IL-4 by $ICOS^{-/-}$ T cells. Enriched T cells from $ICOS^{-/-}$ and WT littermate mice immunized *in vivo* with KLH/CFA were restimulated *in vitro* with KLH (see Methods: 'Analysis of Cytokine Production'). $ICOS^{-/-}$ T cells (black bars) produced lower levels of IL-4 compared with WT controls (white bars), whereas IFN- γ levels were normal. Representative data from one of three similar experiments are shown. **e**, Flow cytometric analysis of intracellular IL-4 and IFN- γ production by $ICOS^{-/-}$ and WT T cells restimulated *in vitro* with KLH. The percentage of IL-4-producing, but not IFN- γ -producing, cells is reduced in $ICOS^{-/-}$ mice. Asterisks indicate statistically significant results.

Methods

Generation of ICOS-deficient mice

We isolated the murine *ICOS* gene from a 129J library using the full-length (800 base pairs) cDNA probe⁵. The targeting vector, which replaced a 2.8-kb genomic fragment with a neomycin resistance (*neo*) cassette in sense orientation relative to *ICOS* transcription, was electroporated into E14 embryonic stem cells (129/Ola). After G418 selection, homologous recombinants were identified by polymerase chain reaction (PCR) using the primer pair 5'-GAG ACT CAT GCT GTG GTT TCA GG-3' and 5'-TTC GCC AAT GAC AAG ACG CTG G-3', and verified by Southern blotting. Chimaeric mice generated from $ICOS^{+/+}$ embryonic stem cell clones were crossed with C57BL/6 females to produce $ICOS^{+/+}$ mice. We assessed germline transmission of the ICOS mutation by PCR and Southern blot analysis of tail DNA. Most of putative exon 2 and all of exons 3 and 4 were deleted, including the putative B7RP1-binding site FDPPPE, homologous to CD28-binding motif MYPPPY²⁶; the entire transmembrane domain; and most of the intracellular domain including the YMFM motif, which has the potential to bind to lipid phosphatidylinositol 3-kinase²⁷.

In vitro stimulations

Lymph-node cells were enriched for T cells by depletion of B220⁺CD11b⁺ cells using a combination of antibody staining (anti-CD11b, Pharmingen), and anti-B220 and anti-rat immunoglobulin magnetic beads (Dyna). ICOS co-stimulation was assessed by stimulation with plate-bound anti-CD3 ϵ ($0.3 \mu\text{g ml}^{-1}$; 2C11, Pharmingen) and anti-human Fc ($12.5 \mu\text{g ml}^{-1}$; Sigma) to bind B7RP1-Fc (ref. 5) ($0.5 \mu\text{g ml}^{-1}$). Background proliferation in response to plate-bound anti-human Fc and soluble B7RP1-Fc was below 200 counts per minute in all instances (Fig. 1d). CD28 co-stimulation was tested using plate-bound anti-CD3 ϵ ($0.3 \mu\text{g ml}^{-1}$) and anti-CD28 ($1 \mu\text{g ml}^{-1}$) mAb. B220-depleted spleen cells were used to test the proliferative response to ConA ($2.5 \mu\text{g ml}^{-1}$) and PMA (10 ng ml^{-1})/ionomycin (100 ng ml^{-1}). Cultures (1×10^5 cells per well) were pulsed with $1 \mu\text{Ci}$ of [^3H]thymidine for 18 h, and [^3H]thymidine incorporation was measured using a Matrix96 direct beta counter (Packard).

Flow cytometric analyses

To assess ICOS expression, cells were stained with B7RP1-Fc (ref. 5) or with a control Fc chimaeric protein, and specific binding was detected using anti-human Fc-fluorescein

isothiocyanate (Jackson Laboratory) as second step reagent. All antibodies and reagents used for surface and intracellular cytofluorimetric analyses were purchased from Pharmingen. We detected cell staining by flow cytometry on FACS Calibur (Becton Dickinson) and analysed it using CellQuest software.

Determination of serum polyclonal isotype-specific immunoglobulin levels

We determined levels of serum polyclonal IgM, IgG1, IgG2a, IgG2b, IgG3 and IgA antibodies using isotype-specific ELISA (Southern Biotechnology). Serial dilutions of sera from experimental mice were prepared and immunoglobulin titres determined by comparing test sample dilution series values with isotype control standard curves using the SOFTmaxPRO (Molecular Devices) ELISA analysis program.

In vivo immunizations

For T-cell-dependent antigens, we immunized mice intraperitoneally (i.p.) with 50 µg KLH (Calbiochem) in CFA (Sigma) or 200 µg NP(18)-Ovalbumin (Biosearch Technologies) in alum. We collected blood from mouse tails at various time points, and we used serum to quantify antigen- and isotype-specific antibodies by ELISA as described below. For secondary responses, KLH-primed mice were challenged with 50 µg KLH in IFA i.p. on day 35 after primary immunization, and bled on day 7 after boosting. NP-OVA sensitized mice were boosted with 100 µg NP(18)-ovalbumin in alum on day 30 and 60 after primary immunization and bled on day 7 after the last boosting. For T-independent antigen analysis, mice were injected i.p. with 25 µg TNP-Ficoll (Biosearch Technologies). We bled tails and prepared serum on days 7, 14, 21 and 28 after immunization.

Measurement of antigen-specific immunoglobulin

Titres of KLH- or NP-specific IgG1, IgG2a and IgM antibodies, or trinitrophenyl (TNP)-specific IgG3 and IgM antibodies, were assessed using sandwich ELISA (Southern Biotechnology Associates) in which ELISA plates were coated with either KLH, NP(23)-bovine serum albumen (BSA) or TNP-BSA at 50 µg ml⁻¹ in PBS or carbonate buffer pH 9.2 (Sigma). We obtained the arbitrary values by using the SOFTmaxPRO (Molecular Devices) ELISA analysis program on the basis of the titration curve of the isotype standard used on each plate. We detected levels of OVA-specific IgE using an antigen-capture (biotinylated OVA) ELISA method as described²⁸. The ELISA was standardized using serum obtained from mice sensitized to OVA according to a conventional i.p. sensitization model²⁹.

Analysis of cytokine production

Mice injected subcutaneously with 50 µg KLH/CFA were killed on day 10 after immunization and draining lymph nodes were isolated. B220- and CD8-depleted lymph node cells (5 × 10⁶ cells per well) were cultured in 12-well plates in the presence of 100 µg ml⁻¹ soluble KLH. We collected culture supernatants after 90 h and levels of IL-4 and IFN-γ were determined by immunoassay (R&D Systems) according to the manufacturer's instructions. For intracellular cytokine determinations, we cultured the stimulated cells for an additional 4 h in the presence of PMA/Iono plus Golgi STOP (Pharmingen) followed by immunostaining and cytofluorimetry.

Statistical analysis of data

The statistical analysis between groups was performed using the exact Jonckheere-Terpstra test implemented on StatXact 4.0.1. This is a non-parametric test that assumes a natural order among the groups, as in the case for ICOS^{+/+} versus ICOS^{+/-} versus ICOS^{-/-}. We adjusted the α-level for multiple comparisons using Dubey's recommendation⁹. P values were significant if they were ≤ 0.01 in the case of serum immunoglobulin basal levels and in the NP-OVA experiment, ≤ 0.02 in the KLH experiment, and ≤ 0.03 in the TNP-Ficoll experiment.

Received 15 September; accepted 2 November 2000.

- McAdam, A. J., Schweitzer, A. N. & Sharpe, A. H. The role of B7 co-stimulation in activation and differentiation of CD4⁺ and CD8⁺ T cells. *Immunol. Rev.* **165**, 231–247 (1998).
- Lenschow, D. J., Walunas, T. L. & Bluestone, J. A. CD28/B7 system of T cell costimulation. *Annu. Rev. Immunol.* **14**, 233–258 (1996).
- Watts, T. H. & DeBenedette, M. A. T cell co-stimulatory molecules other than CD28. *Cur. Opin. Immunol.* **11**, 286–293 (1999).
- Hutloff, A. *et al.* ICOS is an inducible T-cell co-stimulator structurally and functionally related to CD28. *Nature* **397**, 263–266 (1999).
- Yoshinaga, S. K. *et al.* T-cell co-stimulation through B7RP-1 and ICOS. *Nature* **402**, 827–832 (1999).
- Swallow, M. M., Wallin, J. J. & Sha, W. C. B7h, a novel costimulatory homolog of B7.1 and B7.2, is induced by TNFα. *Immunity* **11**, 423–432 (1999).
- Aicher, A. *et al.* Characterization of human inducible costimulator ligand expression and function. *J. Immunol.* **164**, 4689–4696 (2000).
- Coyle, A. J. *et al.* The CD28-related molecule ICOS is required for effective T cell-dependent immune responses. *Immunity* **13**, 95–105 (2000).
- Statistical Methods in the Pharmaceutical Industry* 2nd edn (eds Buncher, C. R. & Tsay, J.-Y.) 517–518 (Marcel Dekker, New York, 1994).
- Grusby, M. J. & Glimcher, L. H. Immune responses in MHC class II-deficient mice. *Annu. Rev. Immunol.* **13**, 417–435 (1995).
- MacLennan, I. C. M. Germinal centers. *Annu. Rev. Immunol.* **12**, 117–139 (1994).
- Nishina, H. *et al.* Impaired CD28-mediated interleukin 2 production and proliferation in stress kinase SAPK/ERK1 kinase (SEK1)/mitogen-activated protein kinase kinase 4 (MKK4)-deficient T lymphocytes. *J. Exp. Med.* **186**, 941–953 (1997).
- Mond, J. J., Lees, A. & Snapper, C. M. T cell-independent antigens type 2. *Annu. Rev. Immunol.* **13**, 655–692 (1995).

- Snapper, C. M. & Paul, W. E. Interferon-γ and B cell stimulatory factor-1 reciprocally regulate Ig isotype production. *Science* **236**, 944–947 (1987).
- Finkelman, F. D. *et al.* Lymphokine control of *in vivo* immunoglobulin isotype selection. *Annu. Rev. Immunol.* **8**, 303–333 (1990).
- Snapper, C. M., Finkelman, F. D. & Paul, W. E. Regulation of IgG1 and IgE production by interleukin 4. *Immunol. Rev.* **102**, 51–75 (1988).
- Mosmann, T. R. & Coffman, R. L. TH1 and TH2 cells: different patterns of lymphokine secretion lead to different functional properties. *Annu. Rev. Immunol.* **7**, 145–173 (1989).
- O'Garra, A. Cytokines induce the development of functionally heterogeneous T helper cell subsets. *Immunity* **8**, 275–283 (1998).
- Kopf, M. *et al.* Inducible costimulator protein (ICOS) controls T helper cell subset polarization after virus and parasite infection. *J. Exp. Med.* **192**, 53–61 (2000).
- Ferguson, S. E., Han, S., Kelsoe, G. & Thompson, C. B. CD28 is required for germinal center formation. *J. Immunol.* **156**, 4576–4581 (1996).
- Xu, J. *et al.* Mice deficient for the CD40 ligand. *Immunity* **1**, 423–431 (1994).
- Renshaw, B. R. *et al.* Humoral immune responses in CD40 ligand-deficient mice. *J. Exp. Med.* **180**, 1889–1900 (1994).
- Shahinian, A. *et al.* Differential T cell costimulatory requirements in CD28-deficient mice. *Science* **261**, 609–612 (1993).
- Grewal, I. S. & Flavell, R. A. The CD40 ligand: at the center of the immune universe? *Immunol. Res.* **16**, 59–70 (1997).
- Metzler, W. J. *et al.* Solution structure of human CTLA-4 and delineation of a CD80/CD86 binding site conserved in CD28. *Nature Struct. Biol.* **4**, 527–531.
- Prasad, K. V. *et al.* T-cell antigen CD28 interacts with the lipid kinase phosphatidylinositol 3-kinase by a cytoplasmic Tyr(P)-Met-Xaa-Met motif. *Proc. Natl Acad. Sci. USA* **91**, 2834–2838 (1994).
- Stampfli, M. R. *et al.* Regulation of allergic mucosal sensitization by interleukin-12 gene transfer to the airway. *Am. J. Respir. Cell. Mol. Biol.* **21**, 317–326 (1999).
- Ohkawara, Y. *et al.* Cytokine and eosinophil responses in the lung, peripheral blood, and bone marrow compartments in a murine model of allergen-induced airways inflammation. *Am. J. Respir. Cell. Mol. Biol.* **16**, 510–520 (1997).
- Kuhn, R., Rajewsky, K. & Muller, W. Generation and analysis of interleukin-4 deficient mice. *Science* **254**, 707–710 (1991).

Acknowledgements

We thank S. Goncharova, J. Haight and C. Smith for technical assistance; M. Pintilie and T. Panzarella for statistical analysis of the data; M. Saunders for scientific editing; K. Bachmaier for critical reading of the manuscript; V. Stambolic and L. Nguyen for comments and advice; and I. Ng for administrative assistance. This work was supported by the Canadian Institutes of Health Research, Amgen Inc., and Canadian Network for Vaccines and Immunotherapeutics of Cancer and Chronic Viral Diseases.

Correspondence and requests for materials should be addressed to T.W.M. (e-mail: tmak@oci.utoronto.ca).

Telomere looping permits gene activation by a downstream UAS in yeast

Derik de Bruin*, Zafar Zaman†, Rachel A. Liberatore* & Mark Ptashne†

* The Rockefeller University, Laboratory of Molecular Genetics & Immunology, 1230 York Avenue, New York, New York 10021, USA

† Molecular Biology Program, Sloan-Kettering Institute for Cancer Research, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

In yeast (*Saccharomyces cerevisiae*), transcriptional activators, such as Gal4 and Gal4-VP16, work ordinarily from sites located in the upstream activating sequence (UAS) positioned about 250 base pairs upstream of the transcription start site¹. In contrast to their behaviour in mammalian cells, however, such activators fail to work when positioned at distances greater than ~600–700 base pairs upstream², or anywhere downstream^{3,4} of the gene. Here we show that, in yeast, a gene bearing an enhancer positioned 1–2 kilobases downstream of the gene is activated if the reporter is linked to a telomere, but not if it is positioned at an internal chromosomal locus. These observations are explained by the finding that yeast telomeres form back-folding, or looped, structures. Because yeast telomeric regions resemble the

heterochromatin found in higher eukaryotes, these findings might also explain why transcription of some higher eukaryotic genes depends on their location in heterochromatin.

The basal level of expression of certain yeast genes—that is, their expression in the absence of an activator—is reduced when they are moved from their ordinary chromosomal location to positions near telomeres. This repression, called telomere position effect (TPE)⁵, is readily overcome by transcriptional activators⁶. TPE depends on the action of the *SIR3* (as well as other) genes⁷, and it has been proposed⁸ that it involves the formation of higher order chromatin structures (for example, DNA looping) at the telomere.

We inserted three reporter gene constructs near the end of the left arm of yeast chromosome VII (Fig. 1). In all three cases, the transcriptional start site of the reporter gene is within 2 kilobases (kb) of the end of the chromosome. In constructs 1A and 1B there is no downstream *UAS_G*. In construct 2, the *UAS_G* is 1.4-kb downstream of the transcriptional start site and is immediately adjacent to the telomere. In construct 3, the orientation of the gene–*UAS_G* pair with respect to the telomere is inverted from that in construct 2, and the *UAS_G* is 1.9-kb downstream of the promoter.

Transcriptional activators elicit transcription of the telomere-linked *URA3* gene bearing a downstream *UAS_G* (Fig. 2). The transcriptional activity of *URA3* is conveniently detected by measuring colony formation on plates containing 5-fluoro-orotic acid (FOA): cells producing Ura3 protein die on media containing this pyrimidine analogue⁹. Figure 2a shows the effects of a series of Gal4 derivatives that differ in their activating potentials as assayed under ordinary conditions (that is, with a reporter bearing an upstream *UAS_G* and positioned away from a telomere)¹⁰. The activators worked on the telomere-linked *URA3* gene as predicted

by the order of their activating potential. Figure 2a also shows that activation was entirely dependent on the downstream *UAS_G* (compare top and bottom panels). Figure 2b shows that a non-classical activator, the Gal4–Gal11 fusion protein, also activated the telomere-linked *URA3* bearing the downstream *UAS_G*, and did so at least as efficiently as wild-type Gal4. Gal4–Gal11 contains the Gal4 DNA-binding domain but lacks the Gal4-activating region; in place of the latter is Gal11, a component of the holoenzyme. This fusion protein works as an efficient activator in yeast, indicating that, in agreement with much other evidence, recruitment of the transcriptional machinery to a promoter can suffice for gene activation¹¹. In all cases, gene activity, as assayed by FOA sensitivity, correlated with increased messenger RNA levels, as assayed by RNA slot-blot analysis (data not shown).

We next measured activation of a *HIS3* gene bearing a downstream *UAS_G* as a function of location near a telomere or at an internal site (Fig. 3). In these experiments, all cells expressed a full-length Gal4 protein. Activity of the *HIS3* gene is conveniently assayed by growth on the inhibitor 3-AT; in this case, in contrast to the assay used for measuring *URA3* activation, growth depends on *HIS3* gene expression. Figure 3 shows that, in wild-type cells, Gal4 activated the telomere-linked *HIS3* in a *UAS_G*-dependent manner, such that cells with the *UAS_G* grew on media that contained 5 mM 3-AT (compare rows 1 and 2). This activation was not observed when the constructs were positioned at an internal locus (compare rows 3 and 4). As expected, Gal4 activated a *HIS3* positioned at an internal locus when the *UAS_G* was inserted just upstream of the gene (row 5).

Gal4-dependent activation of the telomere-linked *HIS3* reporter bearing the downstream *UAS_G* was dependent on the presence of the

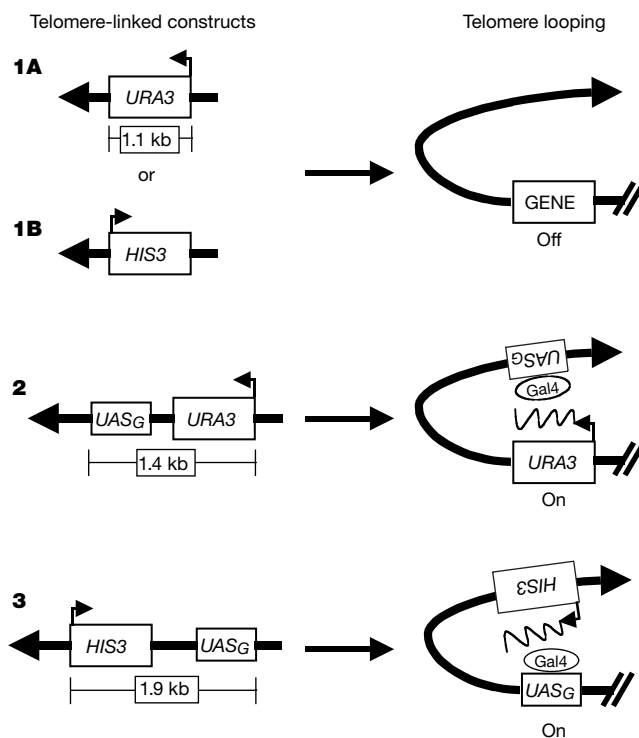


Figure 1 Telomere-linked reporter gene constructs and model of gene activation. Left, constructs placed near the chromosome VII–L telomere. The *C₁₋₃/ATG₁₋₃* telomeric repeat tract is indicated by the heavy arrow. Construct 1A, the *URA3* promoter (indicated by the small left arrow) was located 1.1 kb from the telomere repeat⁵. Construct 1B, the *HIS3* promoter (indicated by the small right arrow) was located immediately adjacent to the telomere repeat and is 300 ± 75 bp from the end of the chromosome²². The *HIS3* gene used in these experiments lacks a *UAS²³*. Construct 2, a native *GAL1, 10 UAS_G* gene containing four Gal4-binding sites, but lacking TATA sequences, was inserted adjacent to

the telomere and immediately 3' of the *URA3* gene in construct 1A²¹. The *UAS_G* in construct 2 was positioned 1.4-kb downstream from the *URA3* promoter. Construct 3, the *UAS_G* was 1.9-kb downstream of the telomere adjacent *HIS3* gene of construct 1B. Note that the orientation of the gene–*UAS_G* pair in construct 3 is inverted with respect to the telomere from that in construct number two. Right, telomere looping on reporter gene activation. Those reporters without the downstream *UAS_G* are not activated in the presence of Gal4 and remain repressed by TPE.

SIR3 protein (Fig. 4). For this experiment we measured *HIS3* mRNA in cells bearing constructs 1B and 3 and grown in media lacking any selective drug. Figure 4 shows that Gal4 activation depended on the downstream *UAS_G* (compare black bars in Fig. 4), and was abolished by mutation of *sir3* (compare bars of construct bearing *UAS_G*). The actual northern blot data is shown on the left. Spot tests similar to those in Fig. 3 yielded results consistent with these results (data not shown).

The results of the following experiments strongly support the suggestion⁸ that yeast telomeres form back-folding loops, and further indicate that telomeric looping is dependent on the activities of the SIR2 and SIR3 proteins. We used chromatin immunoprecipitation (ChIP) experiments (Fig. 5), in which we measured the ability of Gal4, bound to a downstream *UAS_G* (Fig. 1, construct 2), to interact with the *URA3* promoter under three conditions: (1) the reporter was at the telomere in a wild-type cell; (2) the reporter was at the telomere in a *sir3* mutant cell; and (3) the reporter was positioned at an internal (that is, non-telomeric) chromosomal locus. Figure 5 shows (shaded bars) that antibodies to Gal4 precipitated the *URA3* promoter region ~30% as efficiently as they precipitated the downstream *UAS_G* when the reporter was positioned at the telomere in wild-type cells. Figure 5 also shows that little or no Gal4 was detected (white bars) at a sequence within the *URA3* gene. No *URA3* promoter sequences were immunoprecipitated by Gal4 antibodies when the downstream *UAS_G* was absent from the telomere-linked *URA3* (data not shown). The amount of

Gal4 presented at the *URA3* promoter region by the *UAS_G* was decreased about sixfold by mutation of either *sir3* or *sir2* (data not shown), or by positioning the reporter at an internal chromosomal locus.

Our results are consistent with the diagram in Fig. 1, which depicts telomere looping as facilitating interaction of a DNA tethered activator with the transcriptional machinery.

The telomere position effect in yeast is analogous to the position effect variegation (PEV) observed in *Drosophila*, wherein certain euchromatic genes are silenced when placed either near or within heterochromatic regions of the chromosome¹². There are, indeed, parallels between the chromatin structures comprising the telomeric regions of yeast and *Drosophila* heterochromatin. In *Drosophila*, as at yeast telomeres, the heterochromatic regions are less accessible to DNA modifying proteins and nucleases^{13,14}, and in both organisms the associated histones are hypoacetylated^{15,16}. It has been suggested, moreover, that repeated sequences form heterochromatic regions in *Drosophila*¹⁷, implying that such regions comprise folded chromatin structures. Although PEV is typically described as a heritable repression by heterochromatin of a gene that ordinarily resides in the euchromatin, some *Drosophila* genes are expressed only when positioned in heterochromatin^{12,18}.

Our experiments also show that, in yeast, moving an activator-binding site away from a promoter (in this case 1–2-kb downstream) renders activation of that gene dependent on its location within heterochromatin. Perhaps *Drosophila* genes that are

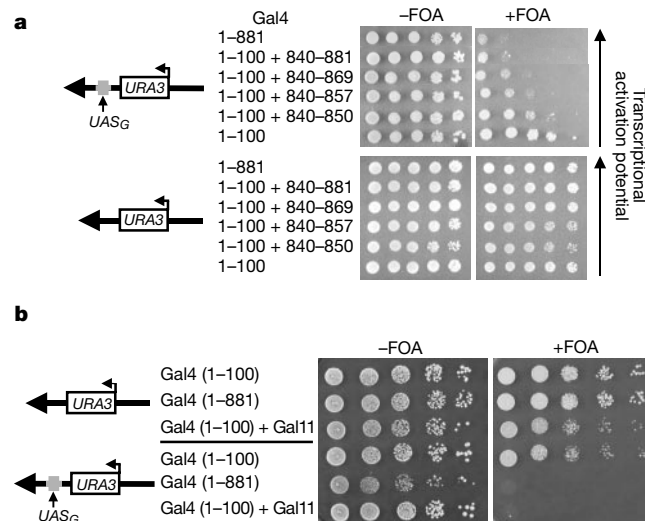


Figure 2 Overcoming TPE by an activator working from the downstream direction. Expression of *URA3* decreases cell viability on FOA plates. **a**, Gal4 derivatives activated a telomere-linked *URA3* gene with a downstream *UAS_G*. **b**, Gal4 (1–881) and a Gal4–Gal11 fusion protein, but not the Gal4 DNA-binding domain alone (Gal4 (1–100)) activated

expression of a telomere-linked *URA3* gene bearing a downstream *UAS_G*. FOA-resistant growth of cells containing the telomere-linked *URA3* gene but lacking the *UAS_G* is a reflection of TPE, because cells bearing the same gene, but placed away from the telomere, fail to grow under these conditions owing to basal *URA3* expression⁵.

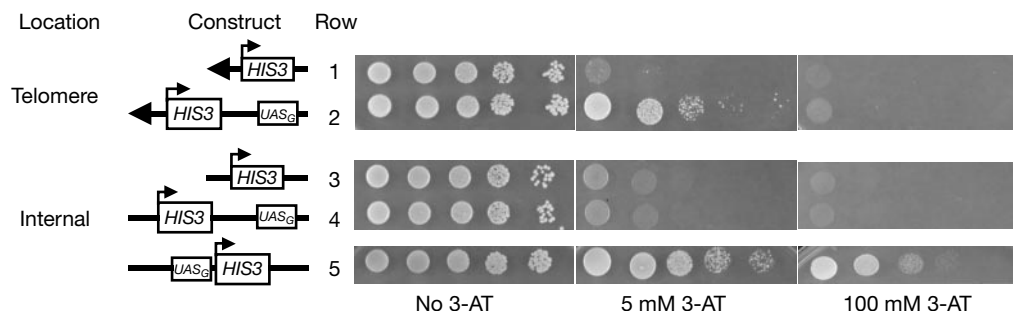


Figure 3 Dependence on telomere linkage for downstream activation. Activity of a telomere-linked *HIS3* gene was spot assayed by measuring colony formation on plates containing the histidine analogue 3-amino-1,2,4-triazol (3-AT): the higher the

concentration of 3-AT, the higher the level of *HIS3* protein required for growth²³. In these experiments, all yeast strains expressed full-length Gal4 (1–881).

expressed only when positioned within heterochromatin (for example, *light*) similarly require looped DNA structures for activation, and such folded structures may be a characteristic of heterochromatin in general^{12,18}. We suggest that the inability of activators to work 'at a distance' in yeast, when the *UAS*-gene construct is located away from a telomere, is accounted for by some limitation on looping. That process, or its equivalent, might be

facilitated in higher eukaryotes by heterochromatic regions and/or by auxiliary factors not present in yeast^{12,17,19}. □

Methods

Plasmids and yeast strains

Yeast strains containing construct 1A and 2 were made by transforming YJOZ (*MAT α gal4 Δ gal80 Δ ura3-52 leu2-3,112 trp1-901 his3 Δ 200 ade2-101 GAL1::lacZ*)²⁰ with plasmids pADH4UCA and pADH4UGT-TATA, respectively, as described²¹. Construct 3 (pGALHISTEL) was made from pYAHISTEL²² by inserting the *UAS_G* (from pADH4UGT-TATA) at the *Sal*I site located 3' of the *HIS3* gene. NLY2 yeast (*MAT α gal4 Δ gal80 Δ his3lys2ura3-52 leu2 trp1*) strains containing construct 1B and 3 were made by transforming with *Pvu*II-digested pYAHISTEL and pGALHISTEL, respectively. To insert the *HIS3* and *HIS3*-*UAS_G* reporters at the internal *URA3* locus, plasmids with (p601*UAS_G*, *HIS3*-*UAS_G*) and without (p601, *HIS3*) the downstream *UAS_G* were digested with *Apa*I before the transformation of NLY2 yeast. Plasmid p601*UAS_G* was made from p601 (ref. 23) by inserting the *UAS_G* (from pADH4UGT-TATA) at the *Sal*I site located 1.9-kb 3' of the *HIS3* gene. The internal control strain with the native *UAS_G* directly upstream of *HIS3*, Y190, has been described²⁰. *SIR3* was disrupted using pNO3 (*sir3::LYS2*) as described²². Other plasmids: Gal4 (pCL1, Gal4 (1–881), *LEU2*, (Clontech)); Gal4 derivatives (1–881, pJR191; 1–100 + 840–881, pJR182; 1–100 + 840–869, pJR192; 1–100 + 840–857, pJR200; 1–100 + 840–850, pJR206; 1–100 (Gal4 DNA-binding domain alone), pJR217)¹⁰; Gal4 (1–100) + Gal11 (799–1081) (pZZ18, *HIS3*).

Spot test assays

For FOA assays, yeast cells were spotted (ten-fold serial dilutions) onto histidine-deficient media (containing 2% raffinose and 1% galactose) either with (+FOA) or without (–FOA) FOA, and then incubated for 4 days at 30 °C before photography⁷. For 3-AT, yeast cells were spotted onto media lacking leucine and histidine (YC-Leu-His), as well as containing the indicated concentration of 3-AT and 2% raffinose/1% galactose. Plates were incubated at 30 °C for 5–10 days before photography.

RNA analysis

We purified total RNA from the *HIS3* or *HIS3*-*UAS_G* reporter-containing wild-type or *SIR3*-disrupted yeast cells grown in media lacking leucine and containing 2% raffinose and 1% galactose. Equivalent RNA samples were run on a 1.1% formaldehyde agarose gel, transferred to a nylon membrane and probed with the 338-bp *Bgl*II fragment of the *HIS3* gene and the 282-bp *Kpn*I and *Pst*I fragment of pYST122 (ref. 24). *HIS3* RNA was quantitated and normalized against *ACT1* RNA.

Chromatin immunoprecipitation

In the yeast strains used for ChIP analysis, the internal *ura3-52* locus was deleted⁵ so that the only *URA3* sequences in the strain came from the reporter construct. To insert the reporter at the internal *LYS2* locus, a *LYS2* containing *Sal*I fragment from pDP6 was inserted into the *Sal*I site of pADH4UGT-TATA²¹, and the resulting plasmid, pXTLYS2, was then cleaved with *Xho*I before transformation. ChIP analyses using antibodies to Gal4 (Santa Cruz Biotechnology) were carried out on 2% raffinose and 1% galactose grown mid-log phase yeast cells expressing *GAL4* as described⁸. Real-time detection and quantitative PCR analysis were done with an annealing temperature of 50 °C using the GeneAmp 5700 Sequence Detection System (PE Biosystems) according to the manufacturer's instructions. Regions A, B and C were amplified using primers pairs UF+UR, PF+PR and F2+F4, respectively. Primer sequences: UF, 5'-GAC-ATT-ATT-ATT-GTT-GGA-AGA-GGA-3'; UR, 5'-GAA-GAG-GAA-AAA-TTG-GC-GT-3'; PF, 5'-GAT-TCG-GTA-ATC-TCC-GAG-CAG-3'; PR, 5'-GTA-GCT-TTC-GAC-ATG-ATT-TAT-CTT-CG-3'; F2, 5'-AGG-CCT-CTA-GGT-TCC-TTT-GTT-AC-3'; F4, 5'-GTC-CCA-AAA-TTT-GTT-TAC-TAA-AAA-C-3'.

Received 26 June; accepted 16 October 2000.

- Guarente, L., Yocum, R. R. & Gifford, P. A GAL10-CYC1 hybrid yeast promoter identifies the GAL4 regulatory region as an upstream site. *Proc. Natl Acad. Sci. USA* **79**, 7410–7414 (1982).
- Keegan, L., Gill, G. & Ptashne, M. Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science* **231**, 699–704 (1986).
- Guarente, L. & Hoar, E. Upstream activation sites of the *CYC1* gene of *Saccharomyces cerevisiae* are active when inverted but not when placed downstream of the "TATA box". *Proc. Natl Acad. Sci. USA* **81**, 7860–7864 (1984).
- Struhl, K. Genetic properties and chromatin structure of the yeast gal regulatory element: an enhancer-like sequence. *Proc. Natl Acad. Sci. USA* **81**, 7865–7869 (1984).
- Gottschling, D. E., Aparicio, O. M., Billington, B. L. & Zakian, V. A. Position effect at *S. cerevisiae* telomeres: reversible repression of Pol II transcription. *Cell* **63**, 751–762 (1990).
- Aparicio, O. M. & Gottschling, D. E. Overcoming telomeric silencing: a trans-activator competes to establish gene expression in a cell cycle-dependent way. *Genes Dev.* **8**, 1133–1146 (1994).
- Aparicio, O. M., Billington, B. L. & Gottschling, D. E. Modifiers of position effect are shared between telomeric and silent mating-type loci in *S. cerevisiae*. *Cell* **66**, 1279–1287 (1991).
- Strahl-Bolsinger, S., Hecht, A., Luo, K. & Grunstein, M. SIR2 and SIR4 interactions differ in core and extended telomeric heterochromatin in yeast. *Genes Dev.* **11**, 83–93 (1997).
- Boeke, J. D., LaCrute, F. & Fink, G. R. A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Mol. Gen. Genet.* **197**, 345–346 (1984).
- Wu, Y., Reece, R. J. & Ptashne, M. Quantitation of putative activator-target affinities predicts transcriptional activating potentials. *EMBO J.* **15**, 3951–3963 (1996).
- Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569–577 (1997).

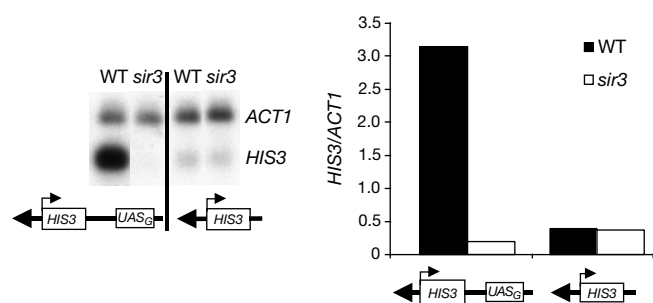


Figure 4 SIR3-dependent downstream activation of a telomere-linked gene as assayed by mRNA levels. A northern blot of total RNA from either wild-type (WT) or *sir3* cells bearing the *HIS3* or *HIS3*-*UAS_G* reporter constructs and expressing full-length Gal4 (1–881) shows that downstream activation of *HIS3* is dependent on both *SIR3* and the *UAS_G*.

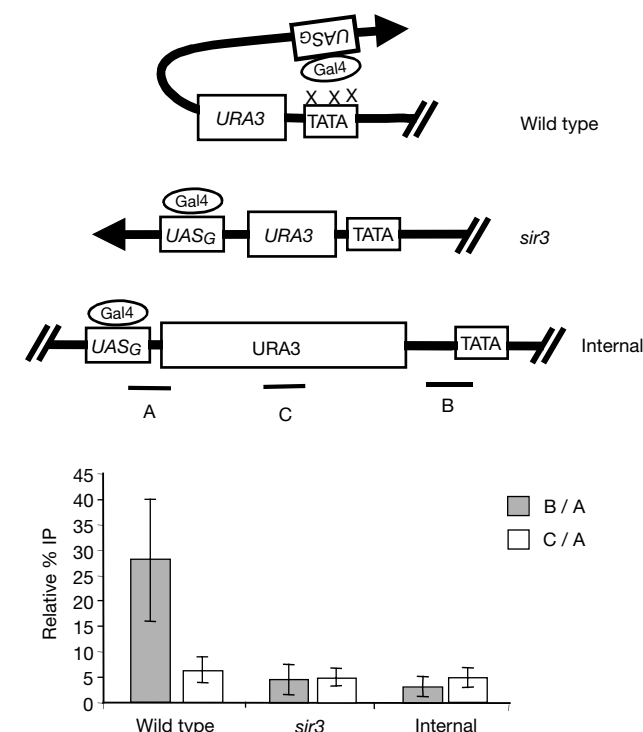


Figure 5 Telomere looping as assayed by chromosomal immunoprecipitation (ChIP). ChIP assays were performed on yeast cells containing the indicated *URA3* reporter constructs located either at the telomere (wild type), at the telomere but in a strain mutated for *sir3* (*sir3*), or at the non-telomeric *LYS2* locus (Internal). Antibodies specific for the Gal4 DNA-binding domain were used to immunoprecipitate chromatin after *in vivo* formaldehyde crosslinking, and the specific DNA regions coupled to Gal4 were analysed by quantitative PCR analysis. DNA fragments corresponding to the downstream *UAS_G* (region A), the *URA3* promoter (region B) and the middle of the *URA3* coding sequence (region C) were amplified by PCR and primer pairs as indicated in Methods. Histogram indicates the percentage DNA immunoprecipitated (IP) at the *URA3* promoter (shaded bars corresponding to region B) and at the middle region of the *URA3* coding sequence (white bars corresponding to region C) relative to that at the downstream *UAS_G* (region A) in each cell type. Plotted values correspond to the mean $\pm$ s.d. of three independent experiments.

12. Wakimoto, B. T. Beyond the nucleosome: epigenetic aspects of position-effect variegation in *Drosophila*. *Cell* **93**, 321–324 (1998).
13. Boivin, A. & Dura, J. M. In vivo chromatin accessibility correlates with gene silencing in *Drosophila*. *Genetics* **150**, 1539–1549 (1998).
14. Gottschling, D. E. Telomere-proximal DNA in *Saccharomyces cerevisiae* is refractory to methyltransferase activity in vivo. *Proc. Natl Acad. Sci. USA* **89**, 4062–4065 (1992).
15. Braunstein, M., Sobel, R. E., Allis, C. D., Turner, B. M. & Broach, J. R. Efficient transcriptional silencing in *Saccharomyces cerevisiae* requires a heterochromatin histone acetylation pattern. *Mol. Cell Biol.* **16**, 4349–4356 (1996).
16. Turner, B. M., Birley, A. J. & Lavender, J. Histone H4 isoforms acetylated at specific lysine residues define individual chromosomes and chromatin domains in *Drosophila* polytene nuclei. *Cell* **69**, 375–384 (1992).
17. Dorer, D. R. & Henikoff, S. Expansions of transgene repeats cause heterochromatin formation and gene silencing in *Drosophila*. *Cell* **77**, 993–1002 (1994).
18. Devlin, R. H., Bingham, B. & Wakimoto, B. T. The organization and expression of the light gene, a heterochromatic gene of *Drosophila melanogaster*. *Genetics* **125**, 129–140 (1990).
19. Dorsett, D. Distant liaisons: long-range enhancer-promoter interactions in *Drosophila*. *Curr. Opin. Genet. Dev.* **9**, 505–514 (1999).
20. Durfee, T. et al. The retinoblastoma protein associates with the protein phosphatase type 1 catalytic subunit. *Genes Dev.* **7**, 555–569 (1993).
21. Sandell, L. L., Gottschling, D. E. & Zakian, V. A. Transcription of a yeast telomere alleviates telomere position effect without affecting chromosome stability. *Proc. Natl Acad. Sci. USA* **91**, 12061–12065 (1994).
22. Bourns, B. D., Alexander, M. K., Smith, A. M. & Zakian, V. A. Sir proteins, Rif proteins, and Cdc13p bind *Saccharomyces* telomeres in vivo. *Mol. Cell Biol.* **18**, 5600–5608 (1998).
23. Alexander, C., Grueneberg, D. A. & Gilman, M. Z. Studying heterologous transcription factors in yeast. *Methods Companion Methods Enzymol.* **5**, 147–155 (1993).
24. Braunstein, M., Rose, A. B., Holmes, S. G., Allis, C. D. & Broach, J. R. Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. *Genes Dev.* **7**, 592–604 (1993).

Acknowledgements

We wish to thank S. Kantrow for her technical assistance and G. Bryant for his help with the quantitative PCR analysis. We are grateful to J. V. Ravetch, Head, Laboratory of Molecular Genetics & Immunology, The Rockefeller University, for his support of D.d.B. and R.A.L. during this work. We also thank T. De Lange and M. Grunstein for comments on the manuscript. M.P. is a Ludwig Foundation Professor. This work was supported in part by NIH and a fellowship from the Norman and Rosita Winston Foundation (Z.Z.).

Correspondence and requests for materials should be addressed to either D.d.B. (e-mail: derik.debruin@ssmb.com) or M.P. (e-mail: m-ptashne@ski.mskcc.org).

Direct observation of DNA rotation during transcription by *Escherichia coli* RNA polymerase

Yoshie Harada^{†‡}, Osamu Ohara[§], Akira Takatsuki^{*}, Hiroyasu Itoh^{†||}, Nobuo Shimamoto[¶] & Kazuhiko Kinoshita Jr^{††}

^{*} Department of Physics, Faculty of Science and Technology, Keio University, Hiyoshi 3-14-1, Kohoku-ku, Yokohama 223-8522, Japan

[†] CREST (Core Research for Evolutional Science and Technology) "Genetic Programming" Team 13, Nogawa 907, Miyamae-ku, Kawasaki 216-0001, Japan

[§] Kazusa DNA Research Institute, Yata 1532-3, Kisarazu 292-0812, Japan

^{||} Tsukuba Research Laboratory, Hamamatsu Photonics KK, Tokodai, Tsukuba 300-2635, Japan

[¶] Structural Biology Center, National Institute of Genetics, Mishima 411-8540, Japan

Helical filaments driven by linear molecular motors are anticipated to rotate around their axis, but rotation consistent with the helical pitch has not been observed. 14S dynein¹ and non-claret disjunctional protein (ncd)² rotated a microtubule more efficiently than expected for its helical pitch, and myosin rotated an actin filament only poorly³. For DNA-based motors such as RNA polymerase, transcription-induced supercoiling of DNA⁴ supports the general picture of tracking along the DNA helix⁵. Here we report direct and real-time optical microscopy

measurements of rotation rate that are consistent with high-fidelity tracking. Single RNA polymerase molecules attached to a glass surface rotated DNA for >100 revolutions around the right-handed screw axis of the double helix with a rotary torque of >5 pN nm. This real-time observation of rotation opens the possibility of resolving individual transcription steps.

Linear movement of a single molecule of RNA polymerase relative to DNA has been observed directly, during transcription^{6–9} and during a one-dimensional diffusional search for a promoter^{8,10,11}. To observe relative rotation between RNA polymerase and DNA, we employed an optical microscopy technique based on the tethered-particle method⁶ (Fig. 1a). A DNA template, 4,971 base pairs (1.7 μ m) long and containing one strong promoter (T7A1), was constructed (Fig. 1b). Transcription was initiated in a bulk solution in the absence of UTP, such that the polymerase stalled at the adenine at the +20 position (A20), before the thymine at +21 position¹². The stalled polymerase was attached to a glass surface (Fig. 1a), and a magnetic bead of diameter 850 nm coated with streptavidin was attached to the downstream end of the DNA where nine nucleotide residues were biotinylated. Then, all four nucleoside triphosphates (NTPs) were added to allow further transcription. In a similar system with a non-magnetic bead, Schafer et al.⁶ observed highly 'processive' (moving over a long distance without detachment) threading of DNA through RNA polymerase,

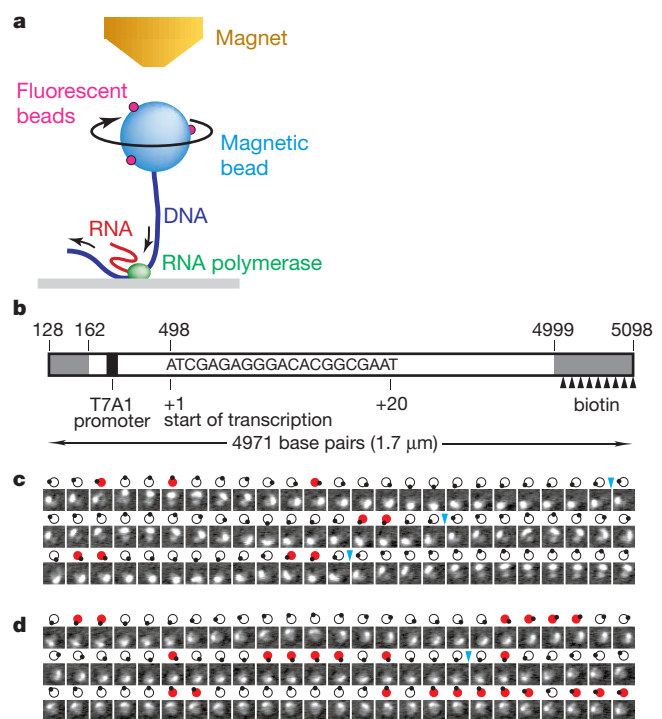


Figure 1 Observation of DNA rotation by RNA polymerase. **a**, Observation system (not to scale). The magnetic bead was pulled upwards by a disk-shaped neodymium magnet, to which a conical iron piece was attached to enhance the magnetic force. Magnetization was vertical and did not prevent bead rotation. Daughter fluorescent beads served as markers of rotation. **b**, The DNA template. Numbers above are from the T7 D111 sequence. Rotation assay started from position +20. The magnetic bead was attached to the nine biotins. Shaded ends denote primers for the polymerase chain reaction. **c**, **d**, Snapshots of rotating beads at 133-ms intervals at NTP concentrations of 50 μ M (**c**) and 2.5 μ M (**d**). Grey part at the centre, a magnetic bead visualized with transmitted light; moving white spot, a daughter fluorescent bead or probably its aggregates. Diagrams show their relative positions. Blue arrowheads indicate completion of a turn. Red diagrams show moments of anticlockwise rotation due to torsional brownian fluctuations of DNA; these were less noticeable at high [NTP], presumably because supercoiling had reduced the effective tether length. Image size, 2.4 $\times$ 2.4 μ m².

[†] Present address: Department of Molecular Physiology, The Tokyo Metropolitan Institute of Medical Science, 18-22, Honkomagome 3-chome, Bunkyo-ku, Tokyo 113-8613, Japan.

visualized as progressive reduction in the range of the brownian motion of the end bead. We also confirmed this. The high processivity, however, does not necessarily imply precise helical tracking. RNA polymerase might occasionally allow DNA to rotate freely within the active site to relieve torsional stress. Or, the polymerase might allow transient, straight backward slippage of DNA without rotation, while imposing helical rotation for forward movement. This would lead to overwinding. For forward movement, RNA polymerase might bring in DNA several base pairs at a time, in which case tracking of the double helix is not required.

We thus attempted to observe rotation directly by decorating the end bead with smaller fluorescent beads. We pulled the magnetic bead upward at ~ 0.1 pN with a magnet for two reasons: to confine the rotation in a horizontal plane, and in the hope of restraining the DNA from supercoiling, which would interfere with torque transmission to the end bead. Up to a few per cent of beads in an observation chamber rotated continuously (Fig. 1c, d), invariably clockwise when viewed from top in Fig. 1a. Threading a right-handed double helix of DNA through RNA polymerase will, in a simple mechanism, produce clockwise rotation. Not all beads rotated, some being stuck on the glass surface; others fluctuated in both directions, presumably being attached to nicked DNA or tethered to an inactive polymerase. Below we report only on those beads that made at least five revolutions in one direction. When NTPs were absent, unidirectional rotation was not observed.

Time courses of rotation of individual beads are shown in Fig. 2. Most curves do not start at time zero, because we had to search over several fields of view before finding a continuously rotating bead. The curves are also shifted upwards to give the impression that all beads started rotating at time zero. In fact we could confirm this only for three cases. Generally the rotation was slower at lower NTP concentration, [NTP], although variations among beads were large. Extremely slow beads, among those examined under the same [NTP], often carried large aggregates of fluorescent beads. Rotation rate also varied with time. Thin parts of the curves in Fig. 2 indicate

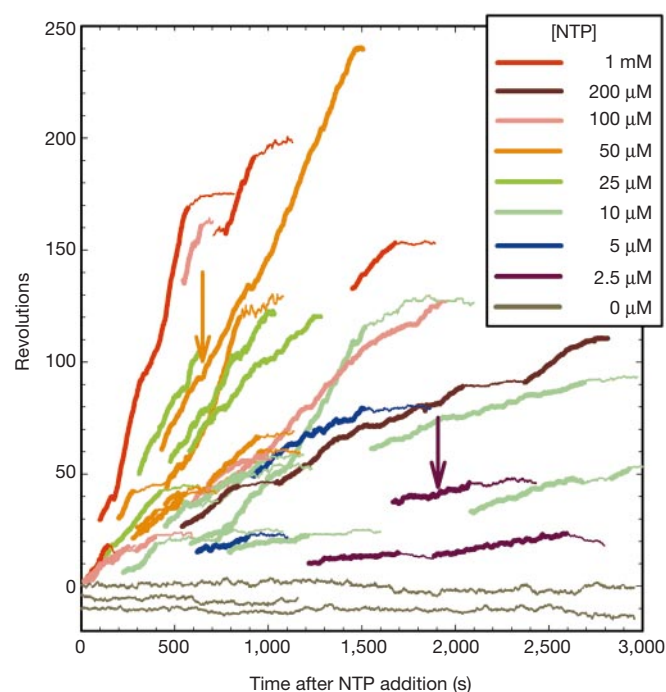


Figure 2 Time courses of bead rotation at various [NTP]. Rotation angles (clockwise positive) were estimated from images as in Fig. 1c, d by centroid analysis¹⁷. Thin lines indicate portions that were excluded in the estimation of rotation rates in Fig. 3. Arrows show the portions shown in Fig. 1c, d.

pauses in the bead rotation; we define a pause as a time interval longer than 50 s during which a bead fluctuated no more than ± 2 revolutions (this was less well defined at low [NTP]). Observation was terminated after a pause, or after complete immobilization for > 30 s, or when the bead tore off and floated into solution. For our DNA template with the full transcript length of $\sim 4,500$ bases, the expected number of revolutions is $4,500/10.4$ (base pairs per turn of DNA¹³); that is, ~ 430 . At least ~ 180 consecutive revolutions have been observed (an orange curve in Fig. 2), suggesting that thousands of base pairs can be transcribed without extensive rotational slippage.

To compare with the rate of rotation, we measured the rate of RNA elongation in solution, and on individual molecules of RNA polymerase on the glass surface⁶ (see Supplementary Information). As summarized in Fig. 3, the two methods gave consistent results (dark versus light green symbols), indicating that surface attachment did not alter the elongation kinetics. The elongation rate divided by 10.4 will be the rotation rate if the polymerase faithfully tracks the DNA helix. Indeed, this seemed to be the case at [NTP] below ~ 20 μM , as seen in red circles in Fig. 3 (also see legend for the purple curve). Although the scatter in rotation data is large, we believe that data showing faster (and longer) rotation are more reliable, because anything that attaches or touches the bead impedes rotation (aggregates of fluorescent beads, DNA, and so on). These data support high-fidelity tracking, but we cannot dismiss the possibility of underwinding by a factor of up to ~ 2 . Extensive overwinding is unlikely.

The apparent saturation of rotation rate in Fig. 3 is explained if the maximal torque of RNA polymerase, Γ_{max} , is already reached at the observed maximal rotation rate of ~ 0.2 revolutions per second (r.p.s.). To rotate a bead of diameter $D = 850$ nm in bulk water at this speed, a torque $\Gamma = 2\pi[0.2 \text{ r.p.s.}] \xi \approx 2.4$ pN nm is required, where $\xi = \pi\eta D^3$ is the rotational frictional drag coefficient and η ($= 10^{-9}$ pN nm⁻² s) is the viscosity of water. The drag is higher

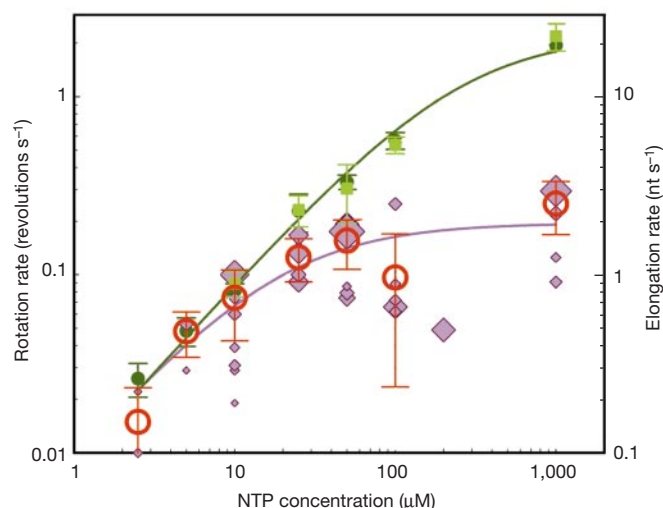


Figure 3 Comparison of rotation and transcription rates. Purple diamonds, rotation rates for individual beads, estimated as the total number of revolutions, N , in the thick portion of a curve in Fig. 2 divided by time; symbol size $\propto N^{1/2}$. Red circles, rotation rates at indicated [NTP] averaged with weights $\propto N$; error bars, root mean square of weighted residuals. Dark green circles, elongation rates in solution; light green squares, elongation rates for individual RNA polymerase molecules on the glass surface (see Supplementary Information). Dark green curve, fit with $V = V_{\text{max}}[\text{NTP}]/([\text{NTP}] + K_m)$ where $V_{\text{max}} = 22$ nucleotides s^{-1} and $K_m = 250$ μM . Purple curve, fit to purple diamonds with rotation rate r given by¹⁷ $1/r = n/V + T_{\text{rot}}$; n/V is the time for elongation reaction where n nucleotides are added per revolution; T_{rot} is the time required to rotate the bead against friction; $n = 8.7 \pm 3.7$ nucleotides per revolution ($n = 10.4$ for precise helical tracking) and $T_{\text{rot}}^{-1} = 0.21 \pm 0.8$ r.p.s.

near the glass surface¹⁴. Pulled at ~ 0.1 pN by the magnet, the bead centre will be ~ 1.1 μm from the surface at the initial DNA length of ~ 1.5 μm (ref. 15). The effective ξ at this height is $\sim 130\%$ of the bulk value, and approaches 300% as the bead comes down towards the surface¹⁴. DNA itself must also rotate, with a similar magnitude of ξ . Altogether, the required torque will be >5 pN nm. Flexible DNA cannot sustain this much torque and collapses by supercoiling, which is expected to occur at the critical torque $\Gamma_c \approx 6$ pN nm under ~ 0.1 pN of tension¹⁵. At high [NTP], beads often rotated steadily without lateral fluctuations, suggesting that extensive supercoiling had made the effective tether length approximately equal to zero. The beads may then touch the surface, further increasing the effective ξ . Presumably, balance between this high drag and Γ_{max} is reached at ~ 0.2 r.p.s. ($= \Gamma_{\text{max}}/2\pi\xi = T_{\text{rot}}^{-1}$ in Fig. 3 legend), resulting in saturation. Γ_{max} must be >5 pN nm, but is unlikely to exceed the 40 pN nm which F_1 -ATPase produces by converting available energy almost entirely to rotation^{16,17}. An alternative explanation for the slow rotation at high [NTP] might be extensive rotary slippage in RNA polymerase at [NTP] > 20 μM . However, simple slippage is unlikely to result in the saturation behaviour, and the transcription obeyed simple Michaelis–Menten kinetics—indicative of an [NTP]-independent mechanism.

In summary, our results indicate that RNA polymerase rotates DNA by tracking its right-handed helix, that the polymerase does so over thousands of base pairs, and that the polymerase can produce >5 pN nm of torque. Whether RNA polymerase rotates around DNA or vice versa is an issue *in vivo*⁵, but our experiment with fixed RNA polymerase cannot answer this problem. Some uncertainty remains in the degree of tracking fidelity, but several beads made many revolutions at the rate commensurate with precise helical tracking. Although the mechanism of tracking is yet unknown, genuine rotary motors may also employ tracking as the rotary mechanism. In the bacterial flagellar motor^{18,19}, the rotor consists of circularly arranged identical subunits. Driving units are also circularly arranged, but one unit suffices to produce efficient rotation¹⁹. A single unit may thus track along the rotor subunits thereby causing rotation. In the F_1 -ATPase in which an asymmetric rotor rotates in 120° steps¹⁷, the tracking mechanism is unlikely.

We expect that the RNA polymerase rotation reported here could provide a means of resolving individual steps of transcription. Transcription of one base will produce a linear translocation of 0.34 nm (ref. 13), which is extremely difficult to resolve. However, accompanying rotation is as much as 35° , and should in principle be detectable. The main problem is torsional brownian motion of DNA, which has to be averaged out. A tag much smaller than the 850-nm bead (and shorter DNA) is needed for averaging within a reasonable time. Determination of the orientation of a tiny tag is feasible, even of a single fluorophore^{3,20}. $\square$

Methods

Materials

Escherichia coli RNA polymerase holoenzyme was purified²¹ and supplemented with excess σ -subunit. The DNA template for transcription was prepared from T7 D111 DNA²², by polymerase chain reaction with an upstream primer for segment 128–162 (in the T7 coordinate) and downstream primer for 4999–5098 (Fig. 1b). The downstream primer had been biotinylated at 9 sites at ~ 10 -base intervals. Before use, the template was treated with T4 DNA ligase at 0.1 mM ATP for 30 min at room temperature to remove nicks. Stalled transcription complex¹² was prepared by incubating 0.3 μM RNA polymerase and 0.3 nM DNA in buffer A (20 mM TrisCl pH 8.0, 20 mM NaCl, 14 mM MgCl₂, 0.1 mM EDTA, 14 mM 2-mercaptoethanol, 1.5% (w/v) glycerol, 20 $\mu\text{g ml}^{-1}$ acetylated BSA, 250 μM ApU dinucleotide) containing 100 μM each of ATP, GTP and CTP for 3 min at 30°C .

Fluorescent, carboxylated microbeads (20 nm, excitation 580 nm, emission 605 nm, Molecular Probes) were amino-derivatized with ethylene diamine in the presence of 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC). The microbeads and Biotin-X cadaverine (Molecular Probes) were conjugated to 850-nm carboxylated magnetic beads (Seradyn) with EDC, and streptavidin was bound to the biotinylated magnetic beads¹¹.

Transcription on the glass surface

A flow chamber¹⁷ was made of two coverslips, which had been sonicated in water, stored in methanol, and dried by setting fire to the methanol. 5 μl of stalled complex in buffer A was mixed with 50 μl of buffer B (20 mM Tris-acetate pH 8.0, 130 mM NaCl, 4 mM MgCl₂, 0.1 mM EDTA, 0.1 mM DTT, 20 $\mu\text{g ml}^{-1}$ acetylated BSA and 80 $\mu\text{g ml}^{-1}$ heparin), infused into the flow chamber, and incubated for 5 min. Further treatments were infusion of 2 mg ml^{-1} α -casein in buffer B for 2 min; washing with buffer B; infusion of beads in buffer C (buffer B minus BSA and heparin, plus 1 mg ml^{-1} α -casein) for 15 min; washing with buffer B plus 1 mM biotin; then washing with buffer B. Transcription was started by infusing NTPs and 0.5% (v/v) 2-mercaptoethanol in buffer B.

Microscopy

Samples were observed at $23 \pm 2^\circ\text{C}$ on an Olympus IX70 inverted microscope with a $100\times$ oil-immersion objective. Magnetic beads were illuminated with a halogen lamp obliquely from above through a ring-shaped optical-fibre assembly. Fluorescent daughter beads were imaged with standard epi-fluorescence optics. Superimposed bright-field and fluorescence images were projected on a silicon-intensified target camera (C2400-08 Hamamatsu Photonics) and recorded on a video tape. A conical magnet was placed above the sample to pull the beads (Fig. 1a). Vertical pulling force was calibrated by tethering the magnetic beads with 16- μm -long λ -phage DNA and measuring the amplitude of brownian motion¹⁵. The force varied among beads and was 0.05–0.2 pN.

Received 25 August; accepted 18 October 2000.

- Vale, R. D. & Toyoshima, Y. Y. Rotation and translocation of microtubules in vitro induced by dyneins from *Tetrahymena cilia*. *Cell* **52**, 459–469 (1988).
- Walker, R. A., Salmon, E. D. & Endow, S. A. The *Drosophila* claret segregation protein is a minus-end directed motor molecule. *Nature* **347**, 780–782 (1990).
- Sase, I., Miyata, H., Ishiwata, S. & Kinoshita, K. Jr Axial rotation of sliding actin filaments revealed by single-fluorophore imaging. *Proc. Natl Acad. Sci. USA* **94**, 5646–5650 (1997).
- Wang, J. C. & Lynch, A. S. Transcription and DNA supercoiling. *Curr. Opin. Genet. Dev.* **3**, 764–768 (1993).
- Cook, P. R. The organization of replication and transcription. *Science* **284**, 1790–1795 (1999).
- Schafer, D. A., Gelles, J., Sheetz, M. P. & Landick, R. Transcription by single molecules of RNA polymerase observed by light microscopy. *Nature* **352**, 444–448 (1991).
- Wang, M. D. *et al.* Force and velocity measured for single molecules of RNA polymerase. *Science* **282**, 902–907 (1999).
- Guthold, M. *et al.* Direct observation of one-dimensional diffusion and transcription by *Escherichia coli* RNA polymerase. *Biophys. J.* **77**, 2284–2294 (1999).
- Davenport, R. J., Wuite, G. J. L., Landick, R. & Bustamante, C. Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase. *Science* **287**, 2497–2500 (2000).
- Kabata, H. *et al.* Visualization of single molecules of RNA polymerase sliding along DNA. *Science* **262**, 1561–1563 (1993).
- Harada, Y. *et al.* Single molecule imaging of RNA polymerase–DNA interactions in real time. *Biophys. J.* **76**, 709–715 (1999).
- Levin, J. R., Krummel, B. & Chamberlin, M. J. Isolation and properties of transcribing ternary complexes of *Escherichia coli* RNA polymerase positioned at single template base. *J. Mol. Biol.* **196**, 85–100 (1987).
- Stryer, L. *Biochemistry* 4th edn (Freeman, New York, 1995).
- Svoboda, K. & Block, S. M. Biological applications of optical tweezers. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 247–285 (1994).
- Strick, T., Allemand, J.-F., Bensimon, D., Lavery, R. & Croquette, V. Phase coexistence in a single DNA molecule. *Physica A* **263**, 392–405 (1999).
- Noji, H., Yasuda, R., Yoshida, M. & Kinoshita, K. Jr Direct observation of the rotation of F_1 -ATPase. *Nature* **386**, 299–302 (1997).
- Yasuda, R., Noji, H., Kinoshita, K. Jr & Yoshida, M. F_1 -ATPase is a highly efficient molecular motor that rotates with discrete 120° steps. *Cell* **93**, 1117–1124 (1998).
- DeRosier, D. J. The turn of the screw: the bacterial flagellar motor. *Cell* **93**, 17–20 (1998).
- Ryu, W. S., Berry, R. M. & Berg, H. C. Torque-generating units of the flagellar motor of *Escherichia coli* have a high duty ratio. *Nature* **403**, 444–447 (2000).
- Adachi, K. *et al.* Stepping rotation of F_1 -ATPase visualized through angle-resolved single-fluorophore imaging. *Proc. Natl Acad. Sci. USA* **97**, 7243–7247 (2000).
- Kubori, T. & Shimamoto, N. A branched pathway in the early stage of transcription by *Escherichia coli* RNA polymerase. *J. Mol. Biol.* **256**, 449–457 (1996).
- Studier, F. W. Gene 0-3 of bacteriophage T7 acts to overcome the DNA restriction system of the host. *J. Mol. Biol.* **94**, 283–295 (1975).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Sasa for help in transcription analysis; A. Ishihama, S. Ishiwata, G. W. Feigenson and members of Team 13 for comments; and H. Umezawa for laboratory management. This work was supported in part by Grants-in-Aid from Ministry of Education, Science, Sports and Culture of Japan, Hayashi Memorial Foundation for Female Natural Scientists, and an Academic Frontier Promotional Project.

Correspondence and requests for materials should be addressed to Y. H. (e-mail: yharada@rinshoken.or.jp).

Lab products in the frame

Apoptosis detection, avoiding western blots and a hand-held video camera.

DC 290 zoom adapter

Kodak www.kodak.com/go/scientific

For a closer look

Specially designed for close-up photography work with the DC 290 digital camera, this new kit includes two +3-dioptre close-up lenses, a threaded lens adapter, a picture card and a macro-framing device to operate with a field of view from 4.25 × 6.25 to 1.25 × 2.25 inches. The kit's Macro-Imaging Script enables automatic camera set-up with a choice of zoom setting, image quality and flash, and options to add annotation and date and time watermarks. Choosing the 'activate at start-up' setting runs the script automatically when the camera is turned on.

Reader Service No. 100

GeneJumper

Invitrogen www.invitrogen.com

Insert transposon primers for DNA sequencing

The GeneJumper kit uses transposon technology to prepare DNA templates for sequencing. Instead of subcloning DNA fragments into vectors with sequencing primer binding sites, this system inserts primer sites randomly into target DNA. Transformation of transposed plasmids generates thousands of sequencing templates, with no need to synthesize custom primers or shotgun clone as in conventional methods.

Reader Service No. 101

Array-Pro Analyzer

Media Cybernetics www.mediacy.com

Microarray analysis software

One of the biggest obstacles in microarray analysis as applied to gene expression has been the extraction of small signal changes in the presence of background noise. This software enables the user to optimize the signal-to-noise ratio by adjusting control parameters and background methods. Standards, controls and replicates enable accurate calibration of signal and background. Computer algorithm automation ensures repeatable results with none of the position changes in spot definition that occur during manual analysis. The software can be combined with GeneMaths for cluster analysis, self-organizing maps and principal component analysis.

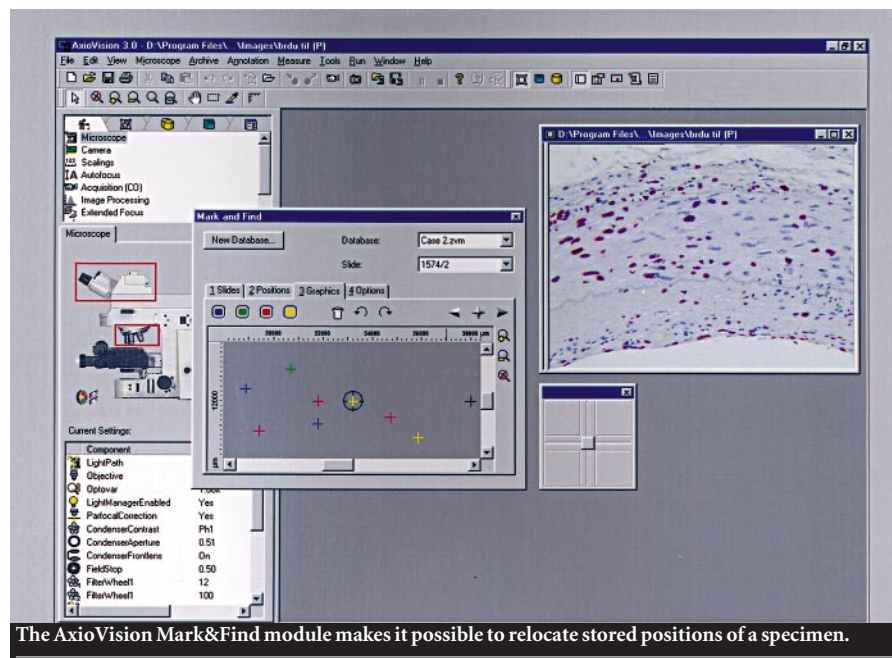
Reader Service No. 102

AxioVision Mark&Find

Zeiss www.zeiss.com

Full marks for this software

Automatic marking of positions on micro-



The AxioVision Mark&Find module makes it possible to relocate stored positions of a specimen.

scope slides and repositioning of marked objects for further processing are the advantages claimed for this software module, which can be used with motorized and encoded stages. A click of the mouse replaces manual recording of mechanical stage coordinates. The software includes a calibration specimen and a database is available for case-related management of several specimens. The display provides an overview of the specimen with position coordinates shown either in coloured marking or as a list. Users can also relocate stored positions of a specimen on other microscopes.

Reader Service No. 103

Apoptosis antibodies

BioVision

Antibodies from A to Z

Apoptosis researchers will find hundreds of products listed in this brochure, including antibodies against caspases (1–11), Bcl-2 family proteins, Apaf-1, DFF40, DFF45, PARP, Fas/Apo1, FADD, TRADD, p53d, cytochrome c, Mcl-1, c-Flip, DcR2, CD40, Erk1/2, Wee1, TNF-R1, ASK1 and IRAK. Caspase assays kits, ready-to-use inhibitors, recombinant caspases (1–10) and a flow cytometric caspase screening kit are also available. BioVision offers annexin V staining kits containing multiple-colour labelling experiments. Other kits include 90-minute DNA ladder detection, cell proliferation, mitochondrial apoptosis detection, ready-to-use apoptosis inducers and *in*

situ DNA fragmentation. The brochure is available online at the company's website.

Reader Service No. 104

FEMLAB 2.0

Comsol www.comsol.com; www.femlab.com

New dimensions in multiphysics modelling

The version 2.0 release of this software allows users to create three-dimensional models on a desktop computer. FEMLAB 2.0 is said to be able to handle virtually any model that can be described by partial differential equations. An enhanced graphical user interface and multi-dimensional visualization tools facilitate rapid design of model geometries, and dynamic solvers permit the dimension of time to be added to models in science and engineering. The software runs on top of MATLAB, the standard application for technical computing.

Reader Service No. 105

GelCode

Pierce www.piercenet.com

Phosphoprotein staining kit

The GelCode kit detects phosphoproteins on SDS-PAGE gels in 3 hours, minimizing the need for western blot detection. The phosphoproteins phosphotyrosine and β -casein are detected in the 40–80- and 80–160-ng per band range respectively in a 4–20% SDS-PAGE gel. Phosphoproteins stain as green to blue-green bands. Each kit includes one positive (phosphotyrosine) and one negative (soybean trypsin inhibitor) control protein and sufficient

new on the market

reagents to stain 10 mini-gels (8×8 cm). Potential applications include dephosphorylation studies.

Reader Service No. 106

Protector XL

Labconco

www.labconco.com

Walk-in laboratory hood

Protector XL hoods come in widths of 8, 10, 12 and 16 feet and in 30- or 36-inch depths and feature by-pass airflow design and a dry-powder, epoxy-coated steel exterior. Chemical-resistant, fire-retardant composite panel liners have a bright white surface to maximize light reflectivity. The horizontal sliding sashes are made of tempered glass with a sash opening for convenient entry. Optional accessories include service fixtures, airflow monitors, distillation grids and a fire extinguisher.

Reader Service No. 107

Microscopy U

Nikon

www.microscopyu.com

Enlarge your view

MicroscopyU is a website offering interactive tutorials on microscopy and photomicrography techniques. It also contains a database of published and unpublished papers, a movie and digital imaging area, featured microscopists, education on optical systems, a microscope museum, information sharing, chat, and an opportunity to enter the International Small World Photomicrography Competition.

Reader Service No. 108

Limtek

Davidson Optronics

www.davidsonoptronics.com

Laser measurement system

The Limtek PC-controlled modular system provides measurements of position, velocity, acceleration, straightness, squareness, parallelism, angles and flatness. It also performs dimensional metrology calibrations, internal multi-axis measurements and high-speed dynamic analysis. Featuring certified laser wavelength, synchronous measurements in up to three axes, an automatic environmental compensation unit, and miniature optics for x/y table positioning, the Limtek comes with comprehensive software packages, allowing a variety of mathematical, statistical and FFT functions to be carried out.

Reader Service No. 109

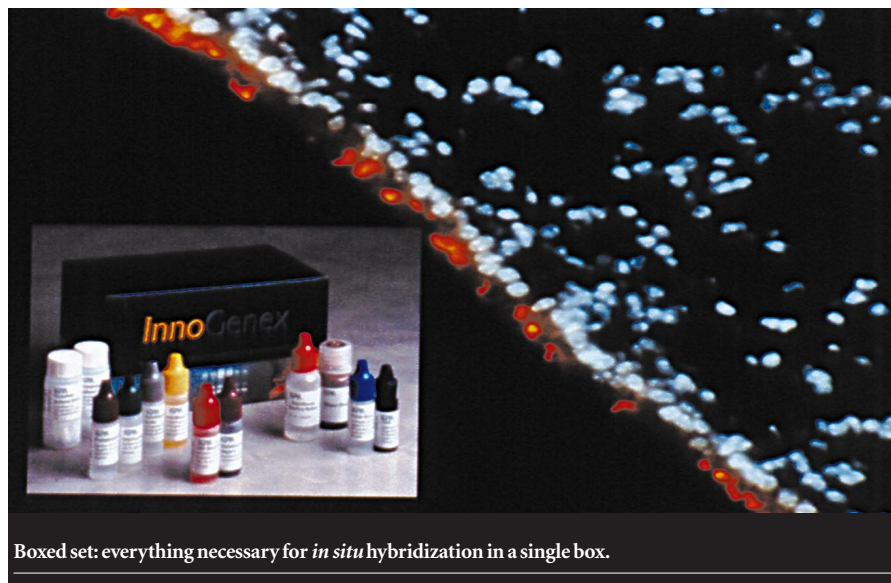
Universal ISH kit

InnoGenex

www.innogenex.com

In situ hybridization for all tissues

This Universal kit for colorimetric detection of fluorescein- and digoxin-labelled probes contains proteinase K solution, RNase block, hybridization buffer, protein block, biotinyl-



Boxed set: everything necessary for *in situ* hybridization in a single box.

lated anti-fluorescein or anti-digoxin antibody, enzyme-streptavidin or fluorochrome conjugate, activation buffer, chromogen/substrate, counterstain, mounting medium and phosphate-buffered saline. All reagents are pre-titrated and supplied in precision dropper bottles. The kit is available with a choice of horseradish peroxidase or alkaline phosphatase conjugate, and AEC or DAB, or BDIP/NBT or Fast Red chromogen/substrate. Each kit contains sufficient reagents for staining 60 slides.

Reader Service No. 110

ThermoCorder

Indigo Systems

www.indigosystems.com

This one's a shoe-in

ThermoCorder is a compact infrared camcorder allowing the integration of the tiny Alpha camera with commercially available Sony camcorders via Sony's intelligent accessory shoe. The resulting infrared images can be viewed on the LCD display or through colour viewfinders. Real-time imagery can

be recorded on DV tape and single images stored on a memory stick. The combined camera and camcorder weigh just 2.6 pounds.

Reader Service No. 111

Gibco BRL brochure

Life Technologies

www.lifetech.com

Essential reading

A new brochure from Gibco BRL outlines the company's range of technologies for research in structural and functional genomics, proteomics and pharmacogenetics. It is aimed at researchers studying gene identification and sequencing, gene localization and positional cloning, expression of proteins, quantification of expression, microarrays, analysis of proteins, protein-protein interactions, and transgenic/knockout models. The brochure is available free of charge online, or by mail.

Reader Service No. 112

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

Custom Peptides & Polyclonal Antibodies

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

Research Genetics

U.S. or Canada 800-533-4363

FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

Seq and ye shall find.....

Complement Genomics offers

Superior DNA Sequencing services!

- ☐ More bases for your money
- ☐ Quality values for each
- ☐ Quick turnaround

Contact us **today** to discuss your requirements

www.complementgenomics.co.uk

TEL: +44 (0) 191 516 6500

FAX: +44 (0) 191 516 6501

READER ENQUIRY NO. 36

READER ENQUIRY NO. 54

© 2001 Macmillan Magazines Ltd

NATURE | VOL 409 | 4 JANUARY 2001 | www.nature.com